University of Alberta

New Solutions to the Measurement Issues in Concept Testing

by

Ling Peng ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
In
Marketing

Faculty of Business

Edmonton, Alberta
Spring 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-29723-0
Our file Notre référence
ISBN: 978-0-494-29723-0

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Although traditional and conjoint forms of concept testing play a pivotal role in the new product development process, they largely ignore data quality issues, as evidence by the traditional reliance on the percent Top 2 Box Scores heuristic. This thesis reconsiders the design and interpretation of the results of concept testing from a measurement theory (Generalizability Theory) perspective. The thesis consists of several research elements, including a survey of new product managers and research consultants to determine the state of concept testing practice, an analysis of four secondary datasets to address important design issues such as sample size and response measures used in concept testing, and a primary study to deal with fundamentals not addressable with the secondary study. This thesis identifies implicit assumptions made about four types of sources (concept related factors, response task factors, situational factors and respondent factors) that can contribute to observed variation in concept testing. Four secondary datasets collected in different contexts provide insight into the assumptions currently made when designing concept testing, and suggest some ways to improve the psychometric quality of concept testing. A three-wave web-based study is conducted to investigate the stability of concept scores. The results show that the three-way interaction among subjects, concepts and occasions is a substantial contributor to variation in concept testing of both major and minor innovations, with the contribution for major innovations even more substantial than minor innovations. Moreover, including occasions as an explicit source of variance lowers the estimates of the generalizability of concept tests. However, the impact of neglecting occasions may vary by purpose of measurement and associated objects of measurement. The primary study also investigates how personality traits in the areas of innovativeness, change seeking and cognitive effort influence concept evaluation scores and whether they could be used to help identify respondents who provide substantially higher quality data in concept testing. The results provide evidence of individual differences for both minor and major innovations in terms of the mean evaluation of concept scores and the generalizability of the data. The latter adds to understanding subject selection issues in concept testing, especially for major innovations.

Acknowledgements

I especially want to thank my supervisor, Professor Adam Finn. His research, insights, criticism and support have contributed greatly to my own thinking, research and writing of this thesis.

A special thanks goes to Professor Todd Rogers, whose encouragement and methodological contributions greatly facilitated this work. Also, I want to acknowledge the considerable benefit I have received from my minor course study in University of Alberta Department of Educational Psychology Center for Research in Applied Measurement and Evaluation (CRAME).

Others who have influenced my work include Professor David Jobson (my multivariate data analysis professor), Dr. Paul Messinger and Dr. Douglas Olsen, who made helpful comments on my thesis proposal. I am also grateful to Professor Roger Calantone from Michigan State University who agreed to serve in my committee.

I would like to thank Gavin Lees and Malcolm Wright, an anonymous Chinese company and a US marketing research firm for their generous help in providing me with the secondary concept testing data sets used in this research. I am also indebted to participants in the marketing series seminar at the University of Alberta School of Business for their helpful comments on an earlier version of parts of this work. Many thanks to all of them.

My primary research was supported by a SSHRC standard research grant to Adam Finn.

Table of Contents

Chapter 1 Introduction 1
- 1.1 Overview of the Thesis 2
- 1.2 Terms and Definitions 3

Chapter 2 Literature Review on Concept Testing 7

Chapter 3 Current Industry Practice 13
- 3.1 The Survey of New Product Managers 14
- 3.2 The Study of Marketing Research Firms 35
- 3.3 Conclusions on the State of Contemporary Practice 37
- 3.4 Limitations 38

Chapter 4 Conceptual Framework and Methodology 40
- 4.1 Generalizability Theory Approach 40
- 4.2 Conceptual Framework and Implicit Assumptions 47

Chapter 5 Secondary Data Studies 51
- 5.1 Academic Data 53
- 5.2 Industry Data for a FMCG 59
- 5.3 Data for Related Innovative Products 60
- 5.4 Discussion and Managerial Implications 63
- 5.5 Conclusions on the Design Issues 65
- 5.6 Limitations of the Secondary Data Studies 66

Chapter 6 Primary Study 69
- 6.1 Theoretical Considerations 69
- 6.2 Method 78
- 6.3 The Generalizability of Concept Testing Over Occasions 83
- 6.4 Individual Differences in Generalizability of Concept Testing 96
- 6.5 Design Issues Revisited 107
- 6.6 Validation Studies 111

Chapter 7 Contributions and Future Work 122
    7.1 Primary Contributions and Conclusions 123
    7.2 Future Work 132

Bibliography 135

Appendices 149
    A-1 Materials for the Survey 149
    A-2 Materials for Pretest 162
    A-3 Pretest Results 169
    A-4 Materials for Main Study 170
    A-5 Output from SPSS GLM-Repeated Measures Analysis 183
    A-6 Confidence Intervals Mat Lab Calculation Code 187
    A-7 Output from SPSS Factor Analysis for the observed trait data 191
    A-8 Output from SPSS Mixed Model Analysis for DSI 193
    A-9 Venn Diagrams for the G Study Designs 195

List of Tables

Table 2-1 Literature Review on Factors Influencing the Evaluative Scores 200

Table 3-1 Market and NPD Focus 18
Table 3-2 Annual Average Number of New Product Introductions 18
Table 3-3 Frequency of Use and Importance of NPD Activities 19
Table 3-4 Models & Methods Frequently Used in NPD Process 21
Table 3-5 Level of Satisfaction with the Various Models & Methods 23
Table 3-6 Annual Average Number of Concept Tests 24
Table 3-7 Comparison of Annual New Product Introductions and Concept Tests 25
Table 3-8 Objectives of the Concept Testing Project 28
Table 3-9 The Design of Tests for Incremental and Radical Concepts 30
Table 3-10 Response Scales Used 31
Table 3-11 Data Collection Methods 32
Table 3-12 Comparisons of Methods for Analyzing the Data 33
Table 3-13 Level of Satisfaction with Predictive Performance 33
Table 3-14 Summary of the Results from the Regression Analysis 34
Table 3-15 Level of satisfaction with various concept testing designs 34

Table 5-1 Details of Four Secondary Datasets 52
Table 5-2 Variance Estimates and Percent of Variance for the Academic Data 54
Table 5-3 Assumed versus Actual Contribution to Variance Made by Sources of Variance in Concept Testing 55
Table 5-4 Designed Reduction of Error Variance When Scaling Concepts 57
Table 5-5 Comparison of G-coefficients When Scaling Concepts 57
Table 5-6 Comparisons of Percent of Variance for Different Evaluation Items 58
Table 5-7 Variance Estimates and Percent of Variance for FMCG Data 59
Table 5-8 Comparisons of Percent of Variance for Different Evaluation Items 60
Table 5-9 Variance Estimates and Percent of Variance for Innovative Test I 61
Table 5-10 Designed Reduction of Error Variance When Scaling Concepts 62
Table 5-11 Variance Estimates and Percentage of Variance for Innovative Test II 62

Table 6-1 Newness and Clearness Scores for Minor and Major Innovations 80
Table 6-2 Variance Component Estimates for Pretest with 10 Selected Concepts 81
Table 6-3 Mean Evaluation Scores by Concepts and Occasions 84
Table 6-4 Variance Component Estimates for Models with Hidden Occasions 87
Table 6-5 Variance Component Estimates Taking Account of Occasions 88
Table 6-6 Comparison of Variance due to CRO 89
Table 6-7 D study Generalizability Coefficients for Concepts 92
Table 6-8 D study Generalizability Coefficients for Concepts by Respondents 91
Table 6-9 Tests of the Hypotheses Made in the Primary Study 95

Table 6-10 Rotated Factor Matrix and Alpha of the Trait Measures 97
Table 6-11 Summaries of the Effects of Traits on Concept Evaluations 99
Table 6-12 Trait Segment Differences in Mean Concept Scores 105
Table 6-13 Trait Segment Differences in G-coefficients 106
Table 6-14 Variance Estimates and Percent of Variance 108
Table 6-15 Comparisons of Percent of Variance for Different Items 110
Table 6-16 Description of Validation Choice Tasks 112
Table 6-17 Concurrent and Predictive Validity at Aggregate Level 115
Table 6-18 Results from Two Binary Logistic Regression Models 118

Table A-1 Newness and Clearness Scores 169

List of Figures

Figure 3-1 Percentage of Total Sales Attributable to New Products 16
Figure 3-2 Relationship between Frequency of use and Critical Importance 20

Figure 4-1 Conceptual Framework 199

Figure 6-1 Profile Plot for Within-subjects Time Effect 86
Figure 6-2 Profile Plot for Within-subjects Time Effect & Newness Effect 86
Figure 6-3 Scree Test for Trait Data 97

Figure A-1 Venn Diagram for the crsi Design in Chapter 4 195
Figure A-2 Venn Diagram for the G Study Design in Academic Data 195
Figure A-3 Venn Diagram for the G Study Design in FMCG Data 196
Figure A-4 Venn Diagram for the G Study Design in Innovative Data I 196
Figure A-5 Venn Diagram for the G Study Design in Innovative Data II 197
Figure A-6 Venn Diagram for the G Study Design in the Pretest 197
Figure A-7 Venn Diagram for the G Study Design in Table 6-5 198
Figure A-8 Venn Diagram for the G Study Design in Table 6-14 198

Chapter 1 Introduction

Concept testing, whether traditional (Moore 1982) or conjoint (Green and Wind 1975), is an important tool used to assess the likely market demand and the best customers to target with potential new products early on in their development (Ozer 1999; Page and Rosenbaum 1992; Crawford and Di Benedetto 2003). Concept testing uses customer response data to pick likely winners from among the many likely losers and to help allocate product development resources. Improving the design and interpretation of the results of concept testing can make an important contribution to improving the efficiency of new product development.

My dissertation research reconsiders the managerial and design aspects of concept testing from a measurement and management decision-making perspective. Traditional concept testing largely ignores data quality issues and relies on simple measurement heuristics, such as percent Top 2 Box Scores on a rating scale. The present research treats the purpose of concept testing as generalizing, with a known error, from a planned set of customer responses to a defined universe of generalization consisting of the conditions under which the products could be marketed once developed. This is done by using generalizability theory, a measurement theory that applies quite directly to concept testing. Generalizability theory can shed light on such issues as the psychometric quality of concept testing data; the optimal design of different types of concept tests; the applicability of concept testing for really innovative products; and whether the potential customers that have been sampled in a concept test include a particularly responsive segment.

When using generalizability theory, scores for any concept are interpreted in light of the managerial decision to be made and the estimated variance components for all facets of variation. This research employs a conceptual framework that identifies four types of sources, namely concept related factors, response task factors, situational factors, and respondent (consumer characteristic) factors. Users of concept testing would like concept related factors to account for a large proportion of variance, and all other factors, including all interactions with concept factors, to be negligible. But, for example, a large variance component due to concepts by respondents indicates that the concepts only appeal to a segment of the tested customer population and less so to other segments of the population.

I first analyze four sets of secondary concept testing data, provided by commercial users, to learn what conclusions can be drawn about the design of concept tests from information about the sources of variation that can be extracted from commercial applications. Issues of direct interest to managers include the sample size needed to reliably scale concepts, whether averaging over multiple items provides more reliable information than a single item, whether the traditional purchase intent item is consistently the best response measure to use in concept testing, and whether much is gained by sampling respondents from multiple locations. Four secondary studies provide insight into the implicit assumptions currently made about the four types of sources, and suggest some ways to improve the psychometric quality of concept testing.

To deal with fundamentals not addressable with secondary data, I then conducted a primary study to incorporate concept testing of (1) minor and major innovations, (2) over multiple test occasions, (3) using consumers, who can be clustered into segments on characteristics assumed to influence their test responses. The main and interaction effects of these three factors need to be well understood in order to use concept testing appropriately, but they have seldom been investigated in the concept testing literature. I collected concept evaluations of ten consumer appliances from members of an online panel (IOCS, the Institute for Online Consumer Studies) on three occasions, approximately a month apart. The primary study examines the temporal stability or generalizability of concept scores over occasions and the individual differences in the generalizability of concept testing.

1.1 Overview of the Thesis

There are seven chapters contained in this thesis. I begin with a review of the literature on concept testing in Chapter 2. Chapter 3 reports the results from a study of new product managers and research consultants to determine the contemporary state of concept testing practice, including the methods and models used, the evidence of their reliability and validity, and practitioners' perception of problems and desired improvements. The two independent sample *t*-test was used to investigate the differences in the approach taken for the testing of incrementally versus radically new concepts.

Chapter 4 presents a conceptual framework for four types of factors that can contribute to the observed variation in concept testing and identifies the implicit assumptions currently being made about these factors in the conduct of concept testing. Since generalizability theory is the major approach used in this thesis, I include an introduction of the G theory approach and how to apply G theory to concept testing in this chapter.

Chapter 5 reconsiders the design of concept testing from a generalizability theory perspective and uses four secondary datasets collected in different concept testing contexts to provide insight into the assumptions currently made when designing concept testing, and suggests some ways to improve the psychometric quality of concept testing.

Chapter 6 describes a web-based primary study of concept testing in which ten innovations are tested on multiple occasions. I investigate the stability of the test results and the importance of occasion as a hidden source of error variance in estimates of the generalizability of concept scores. To investigate the occasion effects in concept testing, repeated measures analysis was used in this portion. I also examine whether a number of personality traits (1) influence concept evaluation scores and (2) can be used to identify respondents who provide substantially higher quality data in concept testing. To investigate possible factors underlying the observed trait data, factor analysis, with principal axis factoring extraction and oblimin transformation, was used. At the end of this chapter I revisit the design issues that were addressed in the secondary data study. Because data were collected on multiple occasions, some validation results can be reported. Binomial logistic regression analysis is used to address the validation questions at the individual level.

Finally, Chapter 7 presents a summary of the conclusions and indicates the areas for future research. Some lessons learned about the design, analysis and interpretation of concept tests that may also apply to other customer research situations, such as prototype/use testing, package testing, and promotion testing, are presented.

1.2 Terms and Definitions

At the end of this chapter, let me define some terms used throughout the thesis. Many terms are used inconsistently in various sources of marketing; I endeavor to be consistent in my usage of these terms throughout the thesis.

Concept testing Concept testing can be "a variety of marketing research-based approaches employed to assess the marketability of a product or service idea prior to its actual development" (Page and Rosenbaum 1992, p. 269). Tauber (1981, p.169) describes the general procedure of concept testing as "consumers are presented with a stimulus (the concept) and measures of reaction are taken which the researcher believes are predictive of the behavioral response, such as later purchase." Moore (1982) identified three types of concept tests, namely concept screening tests, concept generation tests and concept evaluation tests. In my thesis, the focus is on general concept evaluation tests (traditional or conjoint form) where selected consumers evaluate concept stimuli using a set of evaluation items. Traditional concept testing evaluates a concept in isolation while conjoint study uses evaluations of a set of concept designs, with systematic variation in the potential product attributes. Because conjoint tests a designed set of attribute levels, it is a more structured approach to concept optimization and testing (Green and Srinivasan 1990).

Concept testing and premarket forecasting Concept testing and premarket forecasting are two distinct stages in the new product development process. Concept testing aids management in finding the best possible product and marketing execution before the whole package enters the marketplace (a diagnostic focus), while premarket forecasting has a more predictive focus. Concept testing relates to trial behavior and early repeat behavior, but new product success is determined by the adoption level and frequency of purchase. Concept testing cannot measure these two dimensions (Tauber 1975) because consumers with limited exposure to a product concept are not able to predict their own ongoing behavior.

Product Innovativeness (or Product Newness) Product newness is the degree of newness in products (Blythe 1999). Some researchers use definitions of product innovativeness derived from consumer perceptions, while others derive these definitions from the producer's viewpoint (Garcia and Calantone 2002; Calantone, Chan and Cui 2006). This thesis is concerned with product innovativeness from the viewpoint of the consumer, in particular with the ways consumer have of assessing the degree of newness a product possesses. From the consumers' side, product innovativeness refers to the degree of novelty of the product's features / functionality / benefits, degree of change required in consumption pattern and effort required to learn to use and to adopt the new product. The product newness scale used in this thesis is adapted from Lee and O'Connor (2003).

Incremental versus radical innovations A plethora of definitions for innovation types exist in the literature. The terms radical, really new, incremental and discontinuous are used ubiquitously to identify innovations (Garcia and Calantone 2002). In my thesis these terms are used exchangeably. Incremental innovations can easily be defined as products that provide new features, benefits, or improvements to the existing technology in the existing market (Song and Montoya-Weiss 1998). An incremental new product involves the adaption, refinement, and enhancement of existing products and/or production and delivery systems. Examples of products with added features abound in frequently purchased consumer goods (e.g., Tide washing power with bleach and Colgate toothpaste with fluoride and tartar control) (Krieger, et al. 2003). In these kinds of product introductions, the basic product already exists. Features are added to appeal to different buyer segments and extend the basic product's life cycle. Add-on attributes typically entail little risk of buyer alienation and are inexpensive to include. Radical innovations have been defined as innovations that embody a new technology that results in a new market infrastructure (Song and Montoya-Weiss 1998). Radical innovations attempt to break the mold, in terms of function, design, performance, novelty, or other characteristics that separate them from the pack (Urban, et al. 1996). Hybrid cars, high-definition TV, DVD players, PDAs, designer jeans, digital cameras, and Internet auctions are illustrations of such products and services.

Minor versus major innovations My thesis takes a consumer's perspective toward product newness and uses the term minor and major innovations to recognize the fact that there are degrees of newness. The measurement of product newness is based on a continuum of newness between incremental and radical that focuses on perception of newness to consumers. Major innovations are more innovative products that have a high degree of newness and minor innovations are less innovative products that sit close to the opposite end of the continuum.

Generalizability theory and classical test theory Generalizability theory is a statistical theory for evaluating the dependability of measurements (Cronbach, et al. 1972). It accounts for the multi-faceted nature of error and recognizes that measures are used for different decision-making purposes. The basic terms of Generalizability theory are introduced in Chapter 4. Classical test theory is a measurement theory that postulates that an observed measurement can be decomposed into a "true" score and a single undifferentiated random error term (Brennan 2001). As such, any single application of the classical test theory model cannot clearly differentiate among multiple sources of error.

Accuracy of concept evaluation data Accurate concept evaluation data means responses to new product concepts that truly and reliably reflect real differences in consumers' evaluations of the products. In Generalizability theory, accuracy (data quality) is determined by a G coefficient for particular objects of measurement. Higher quality (higher accuracy, higher generalizability) means a G coefficient for the objects that is closer to one.

Chapter 2 Literature Review on Concept Testing

Webster's says a concept is an idea or an abstract notion. Crawford and Di Benedetto (2003) define a complete new product concept as a statement describing the novel features of the innovation relative to existing alternatives in the market place. In a broad sense, a concept is no longer just a simple listing of proposed product attributes, a paragraph description, a pictorial representation, such as a drawing or storyboard or physical mockup, or prototype (Batsell and Wind 1980); it can also be a virtual version of the product (Dahan and Srinivasan 2000). Finn (1985) reviews the typologies of product meanings or attributes of products and concludes that there are physical characteristics, benefits, and imagery aspects to the meaning conveyed to consumers by a new product concept or a product in the marketplace. Therefore concept stimuli can be presented either in terms of physical characteristics or in terms of a limited number of key benefits and physical characteristics, leaving subjects to infer the imagery and any other benefits and physical characteristics.

Concept testing plays a pivotal role in the new product development process. Concept testing refers to estimating "customer reactions to a product idea before committing substantial funds to it" (Moore 1982, p. 279). It can be "a variety of marketing research-based approaches employed to assess the marketability of a product or service idea prior to its actual development" (Page and Rosenbaum 1992, p. 269). Tauber (1981, p. 169) describes the general procedure of concept testing as "consumers are presented with a stimulus (the concept) and measures of reaction are taken which the researcher believes are predictive of the behavioral response, such as later purchase." A typical concept test presents selected respondents with a statement describing a new product idea and asks them to respond to questions, such as liking and purchase intent, which are indicative of its market potential.

Borrowing from Moore (1982) and Anschuetz (1996), there are three types of concept tests, namely concept screening tests, concept generation tests and concept evaluation tests. *Screening tests* seek to reduce a large number of ideas to a more manageable set, where concept statements represent only the core idea. *Concept generation tests* are used to develop a statement that fully describes the product – its physical characteristics and sensory associations and its benefits to the consumer; the concept statement presented should be as clear and meaningful as possible. *Concept evaluation tests* include general concept evaluation, positioning, and concept/product tests, which measure a large number of consumer responses to the concept statement in a more quantitative manner. In this research the focus is on general concept evaluation tests where selected consumers evaluate concept stimuli using a set of evaluation items.

The primary purpose of concept testing has long been viewed as the elimination of poor concepts (Nowland 1947), but it can also be viewed as the identification of extremely attractive opportunities (Dahan and Mendelson 2001). According to the review of the concept testing completed by Tauber (1981, p. 169), "the general purpose of concept testing is to screen out losers at the idea stage through some form of consumer evaluation." Page and Rosenbaum (1992, p. 269) identify providing early feedback from the market about the perceived attractiveness of a proposed new product before its expensive phases of development as the primary objective of concept testing. Crawford and Di Benedetto (2003) note that the key purposes of concept testing are (1) to eliminate poor concepts; (2) to generate a crude estimate of the sales or trial rate; and (3) to develop further the original idea. They observed that managers initially combine the number of people who indicate, "Definitely would buy" or "probably would buy" on the purchase intent question into an aggregate score called the top-two-box score. Marketers then use past experience or an industry rule of thumb (Bell 1994) to convert this score into a prediction of actual purchase or sales. The Booz Allen Sales Estimating System (BASES 2005) group uses concept-test data to evaluate a concept's sales potential and compare its volume potential relative to other concepts. To generate these sales estimates, BASES integrates responses to the purchase intent question with responses to other measures, such as value perception, transaction size and purchase frequency. The resulting sales forecast can be adjusted further by incorporating other factors, such as the expected type of the product, the level of supporting marketing efforts and the idiosyncratic regional characteristics. Concept testing may also help define the highest potential segment of the general target market and provide diagnostic information that can be used to improve the concept and the likelihood of successful introduction. Thus, concept testing often has multiple objectives, not just the scaling of concepts. It is also about the interaction between concepts and their potential attributes and the interaction between concepts and segments of respondents. Thus concept tests need to provide adequate data quality for multiple objects of measurement. A data collection design that adequately scales concepts is not guaranteed to adequately scale other objects of measurement.

A variety of specific methods have been developed for consumer testing of new product concepts (see, for example, Wind 1973; Page and Rosenbaum 1992; Dahan and Srinivasan 2000). The implementation of concept testing requires decisions on numerous design issues, such as what is being tested, who should do the test, and when and how to do the test (Moore 1982). Finn (1985) categorizes these design issues as stimulus-related and subject selection-related. Crawford and Di Benedetto (2003) note the need for decisions about such non-concept factors as the format of the concept, whether or not to offer competitive information, whether or not to put a price in the concept statement, the definition of the respondent group, and the selection of the response situation. Other concerns in concept testing include blind versus brand name testing, immediate assessment versus waiting until after prolonged use in a product test, and the applicability of concept testing for major or discontinuous innovations. Most recently, Klink and Athaide (2006) summarize the three basic design decisions inherent to concept testing, namely stimuli design, respondent selection and response measurement.

Concept testing is not without its critics. Several studies have cast legitimate doubt on the predictive performance of concept testing (Taylor, Houlahan and Gabriel 1975; Morrison 1979; Kalwani and Silk 1982) and on its rightful role in the new product development process (Tauber 1975; Moore 1982; Page and Rosenbaum 1992; Duke 1994; Crawford and Di Benedetto 2003). My literature review reveals that there is information available about the reliability and validity of concept testing, particularly as it relates to the relationship between purchase intent and trial. Taylor, Houlahan and Gabriel (1975) used a field test to conclude that (1) attitudes towards the product during concept test did not have an effect on subsequent searching in the market; and (2) there is a positive relationship between buying interest and purchase behavior, but the purchase intent question only predicts behavior correctly about one time in three. Morrison (1979) proposed and tested a model of the linkage between intentions and purchase for durables, which, among other things, represents several threats to the predictive validity of intention measures. Kalwani and Silk (1982) reported some further analyses and applications of Morrison's model of the predictive relationship between measures of intentions and subsequent purchasing behavior. They found that the presence of substantial components of random and/or systematic errors (e.g., response style) affects adversely their predictive accuracy as well as their ability to discriminate among alternative stimuli, and that the nature and sources of systematic error present in intentions rating are different for durable goods and branded package goods. Mahajan and Wind (1992) found only 19 percent of users of concept tests were highly satisfied. The major reported shortcoming was forecast inaccuracy, cited by 62 percent of users. There are notorious examples of concept testing forecasting failures. For example, the RCA Selectavision VideoDisc system was a technological success when introduced in 1981, but was an expensive failure, falling far short of the volume suggested in market research conducted in the 1970s (Graham 1986). Poor predictive performance could result from low quality measurement practice, changes made to the concept, its positioning, its physical product form between the time of the concept test and market introduction, or from changes in the legal or social environment (Duke 1994; Moore 1982). More recently, Klink and Athaide (2006) use diffusion theory to identify potential sources of concept-test error, and show how results of conventional concept testing can be sensitive to respondents' adoption orientation and the response measure used. Given this concern about forecast accuracy, improving the design and interpretation of the results of concept testing can make an important contribution to improving the efficiency of new product development.

There has been long running debate over the applicability of concept testing techniques to innovative new products. Tauber (1974) argues that more radical innovations should not be subject to traditional concept testing because of the inherent bias against such innovations. Duke (1994) suggests that concepts tests are not very successful for radically new products, where customers have no frame of reference. It is generally accepted that concept testing does a better job of predicting trial for concepts that are not radically different from products already on the market than it does for radical innovations because consumer attitudes, upon first exposure to discontinuous innovations, are not good predictors of what their actions will be after a prolonged exposure. Therefore, some research has attempted to accelerate consumer learning for really new products by providing sources of information that may be available during adoption. Information acceleration, which provides access to the potential benefits prior to measuring a reaction, has been proposed as a way to improve preference measurement. For example, Urban, Weinberg and Hauser (1996) attempt to build consumer knowledge by means of a simulated product search and an evaluation exercise. Hoeffler (2003) examines techniques for incorporating mental simulation and analogies into an existing preference measurement technique and shows how these methods enhance or hinder predictive accuracy.

(Insert Table 2-1 about here)

Table 2-1 provides a structured summary of the empirical research on concept testing. As shown in the Table, empirical studies of concept testing have generally tested design factors in isolation (e.g., "concept formulation" in Lees and Wright 2004; "consumer expertise" in Schoormans, Ortt and de Bont 1995; "competitive-set information" in Miller, Bruvold and Kernan 1987). Moreover they employ fixed effects methods that do not address the generalizability of a particular study's findings. Thus, little is known about the contribution of factors such as concepts, respondents, response items, test occasions, and their important interactions (e.g., concepts by respondents) make to the evaluative responses obtained in concept testing.

This literature review of concept testing reveals first that measurement issues have been largely neglected. While researchers have examined how well concept tests predict new product success (e.g., Taylor, Houlahan and Gabriel 1975; Kalwani and Silk 1982), rarely have they addressed how to ensure a concept test provides a required degree of psychometric quality. New product evaluation research has relied on the Classical Test Theory (CTT) based measurement approach, which treats error as undifferentiated, despite the extensive literature identifying the existence of various types of measurement error (for a review, see Viswanathan 2005). Second, less attention has been given to respondent selection, sample size determination, and the reliability and validity of conventional evaluation items. Third, concept testing has been thought of as a specific research technique, rather than as a measurement process designed to facilitate decision-making. Though some research has investigated specific factors that contribute to test outcomes, little has been done to systematically quantify these factors. The existing literature contains very little discussion of how to quantify factors influencing the testing results or what action to take to ensure a concept test provides a required degree of accuracy. Traditionally, concept-testing researchers have been concerned about identifying the optimal standardized testing condition for controlling non-product sources of variation.

My dissertation research uses generalizability theory to re-examine new product concept testing from a measurement and management decision making perspective.

Chapter 3 Current Industry Practice

New product concept testing plays a pivotal role in the early screening of new product ideas and is considered one of the most critical steps in the new product development process (Ozer 1999). The Mahajan and Wind (1992) survey of Fortune 500 firms (hereafter M&W) found that 87%[1] of the firms used new product concept screening and 72% used customer tests of products. Moreover, concept screening was rated the most critical of the new product development activities. Concept tests, used by 26% of the respondents, were the third most widely used of 24 product development models or methods, well ahead of conjoint analysis (15%) and trailing only focus groups (68%) and limited rollout (42%). However, only 19% of those using concept tests expressed high satisfaction with them. The major reported shortcoming was forecast inaccuracy, cited by 62% of the users.

New product developers now have even more difficult challenges and exciting opportunities. Marketers, who face fierce competition, market and product globalization, and more demanding and sophisticated consumers, are more dependent on new product success to secure close and lasting relationships with their customers. Internet based and virtual concept testing have been heavily promoted research services, targeted at product developers (Dahan and Srinivasan 2000; Dahan and Hauser 2002). Thus, it is timely for developers, suppliers and users of concept testing to assess the role it now plays in the new product development process. Towards this objective, I conducted a study in fall 2005 to determine the current state of concept testing practice.

The purposes of the study were to (1) better understand current concept testing practice and its role in the new product development process; (2) identify the relationship, if any, between current concept testing design and manager perceptions of its effectiveness; and (3) determine what evidence product managers or research consultants have for the reliability and validity of current concept testing. To place the concept testing results in context, the study also revisited some of the more general product development issues examined by M&W. The findings should help in understanding current concept testing practice, including which methods/models are used, what is known about their reliability and validity, and any perceived problems and desired improvements. The results should not only have implications for product developers, research suppliers and users of concept testing, they could also help identify those issues most in need of follow-up research.

---

[1] Here I cited the figures directly from the Mahajan and Wind (1992) survey. To facilitate the comparison between their survey and my survey results, percentages are consistently reported to zero decimal place in this chapter. In other chapters, percents are reported to one decimal place as usual.

3.1 The Survey of New Product Managers

A survey of new product managers was conducted to collect general information on new product development and concept testing over the last three years. In addition, it sought detailed information on the organization's most recent traditional or conjoint concept testing projects where the outcome had been determined and a resulting management decision had been made. The survey was also designed to investigate any differences in the approach taken for incremental versus radical new concepts.

The survey was structured to investigate some common suppositions about the concept testing of radical versus incremental new products. First, radical product innovations require far larger investments and have higher failure rates than less innovative, incremental new products, such as product improvements, repositionings, or extensions (Cooper 2000; Golder and Tellis 1993). Thus, one would expect new product managers to want their concept testing to be as rigorous as possible for radical product innovations (Supposition 1). If so, when testing radical new products, firms should more carefully screen respondents on criteria beyond simple product use or demographics (Supposition 1a), employ larger samples of respondents (Supposition 1b), collect more detailed evaluation information from each respondent (Supposition 1c), and analyze the data collected using more sophisticated methods (Supposition 1d).

Second, it has long been assumed that however rigorous a procedure is, concept testing will predict trial more accurately for minor innovations than it will for radical innovations (Tauber 1974; Gourville 2005). If that were the case, one would expect that after controlling for differences in the testing procedures they employ, new product managers who have tested innovative concepts would be less satisfied with the predictive performance of their concept testing than managers who have tested incremental new products (Supposition 2).

The suppositions about the concept testing of radical versus incremental new products are summarized as follows:

Supposition 1: *New product managers want their concept testing to be as rigorous as possible for radical product innovations.*

Supposition 1a: *When testing radical new products, firms should more carefully screen respondents on criteria beyond simple product use or demographics.*

Supposition 1b: *When testing radical new products, firms should employ larger samples of respondents.*

Supposition 1c: *When testing radical new products, firms should collect more detailed evaluation information from each respondent.*

Supposition 1d: *When testing radical new products, firms should analyze the data collected using more sophisticated methods.*

Supposition 2: *New product managers who have tested innovative concepts would be less satisfied with the predictive performance of their concept testing than managers who have tested incremental new products.*

Data Collection Approach

To obtain detailed information on concept testing methods, the survey was targeted at managers who are responsible for new product development, as they were considered the best source of informed responses. New product managers are knowledgeable about the various models and methods, and most have the responsibility for designing and executing new product concept tests. To reach such respondents, 400 members of a New Product Development (hereafter NPD) association and who identified themselves as product managers were emailed a description of the purpose of the study. The email asked these managers to participate themselves if they were responsible for NPD or to identify a manager from their firm who was most qualified to participate in the study. A total of 111 new product managers agreed to participate. Each respondent was emailed a self-administrative questionnaire in fall of 2005 (Please see materials for the survey in Appendix 1). Combined with a follow-up to non-respondents, this procedure generated

51 usable responses. The firms represented in the sample of new product managers were 20% industrial, 14% service, 18% package goods, 16% durables and 33% others.

To highlight current practice the results reported below are compared with those obtained by M&W where possible. However, it is important to recognize that the differences could not only reflect changes over time, but could also be due to differences in the sampling frame or the fact that managers more often engaged in concept testing were most likely to respond to my survey.

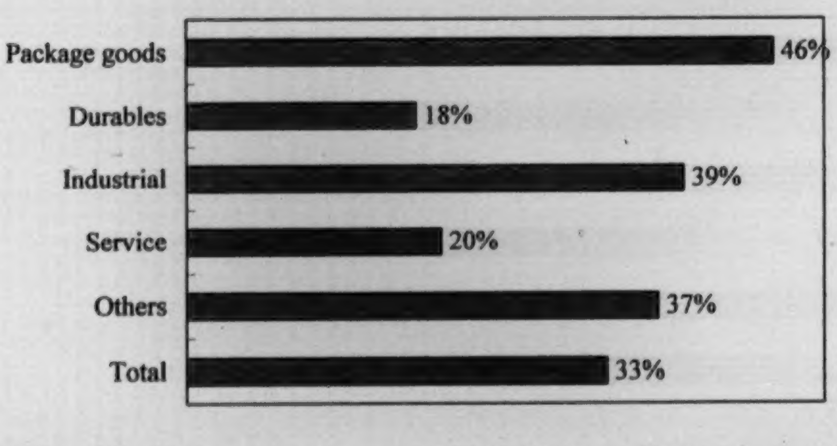
Product Development Characteristics of Responding Firms

Some characteristics of the responding firms are highlighted in Figure 3-1, Table 3-1 and Table 3-2. The following summary is warranted from these exhibits:

1. The percent of total company sales attributable to new products that are less than 3 years old averaged 33%, but ranged as high as 46% for package goods companies and as low as 18% for durables companies (Figure 3-1).

Figure 3-1

Percentage of Total Sales Attributable to New Products



2. The market focus of the company was global (as opposed to domestic) for 73% of firms, while the more specific new product development focus was global for 71% of the companies (Table 3-1). These results suggest new product development is far more global in orientation for these companies now than was reported by M&W.

3. The highest average number of new product introductions per year occurs for durables (29), followed by package goods (19), industrial (12), others (7) and service (5).

Only a small proportion of new introductions are innovative products, particularly for durables and package goods firms (Table 3-2).

(Insert Table 3-1 about here)

(Insert Table 3-2 about here)

Thus the responding companies are actively involved in introducing new products to meet the needs of the global market. They are also introducing more new products than the companies included in the 1992 M&W study. M&W reported an average of 7 new products per year were introduced from 1987-1989 while I report 13 for 2002-2004. They reported that new products made up 25% of total company sales while I report 33%.

The Study Results

New Product Development Activities in the Last Three Years

M&W identified 10 NPD activities that were undertaken by most companies for at least one product in 1987-1989. However the frequency with which an activity was performed for all new products and the importance of an activity varied more widely. Hence data were collected on the latter two issues.

As shown in Table 3-3, product development and business/financial analyses are, respectively, performed for all new products by 55% and 47% of the companies. These two activities are also most often considered of critical importance in the development of a new product. On the other hand, formal new product concept screening, detailed market study for market identification, positioning and strategy, and market test/trial sell are performed infrequently. Not surprisingly, there is a strong relationship between frequency of use and the relative importance of an activity in the new product development process (Figure 3-2).

Comparing the current results with those from M&W, it is interesting to note that business/financial analysis (62% cf. 55%) and market launch planning (50% cf. 41%) are activities increasingly recognized as of critical importance.

Table 3-1
Market and NPD Focus

		Company market focus Domestic (%)	Global (%)	
New Product Focus	Domestic (%)	24 (56)	6 (9)	29 (65)
	Global (%)	4 (4)	67 (31)	71 (35)
		27 (60)	73 (40)	

The corresponding results from Mahajan and Wind (1992) are given in parentheses.

Table 3-2
Annual Average Number of New Product Introductions

Company	2002		2003		2004		Annual Average	
type	All new products	Innovative products	All new products	Innovative products	All new products	Innovative products	All new products	Innovative products
Package	28	0.3	17	0.3	14	1.4	19	0.8
goods	(60)	(0.5)	(31)	(0.5)	(29)	(1.2)		
Durables	41	2.0	24	1.3	27	2.3	29	1.8
	(44)	(2.2)	(35)	(1.8)	(36)	(3.5)		
Industrial	12	2	12	2.4	13	3.3	12	2.6
	(16)	(3.4)	(15)	(3.1)	(15)	(3.3)		
Service	5	0.8	4	1	5	1.2	5	1
	(3)	(1.1)	(2)	(0.9)	(3)	(0.4)		
Others	5	1.1	6	1.6	8	2.3	7	1.7
	(4)	(1.6)	(5)	(2.6)	(7)	(2.8)		
Total	14	1.2	12	1.5	13	2.2	13	1.7
	(30)	(2.0)	(20)	(2.3)	(21)	(2.6)		

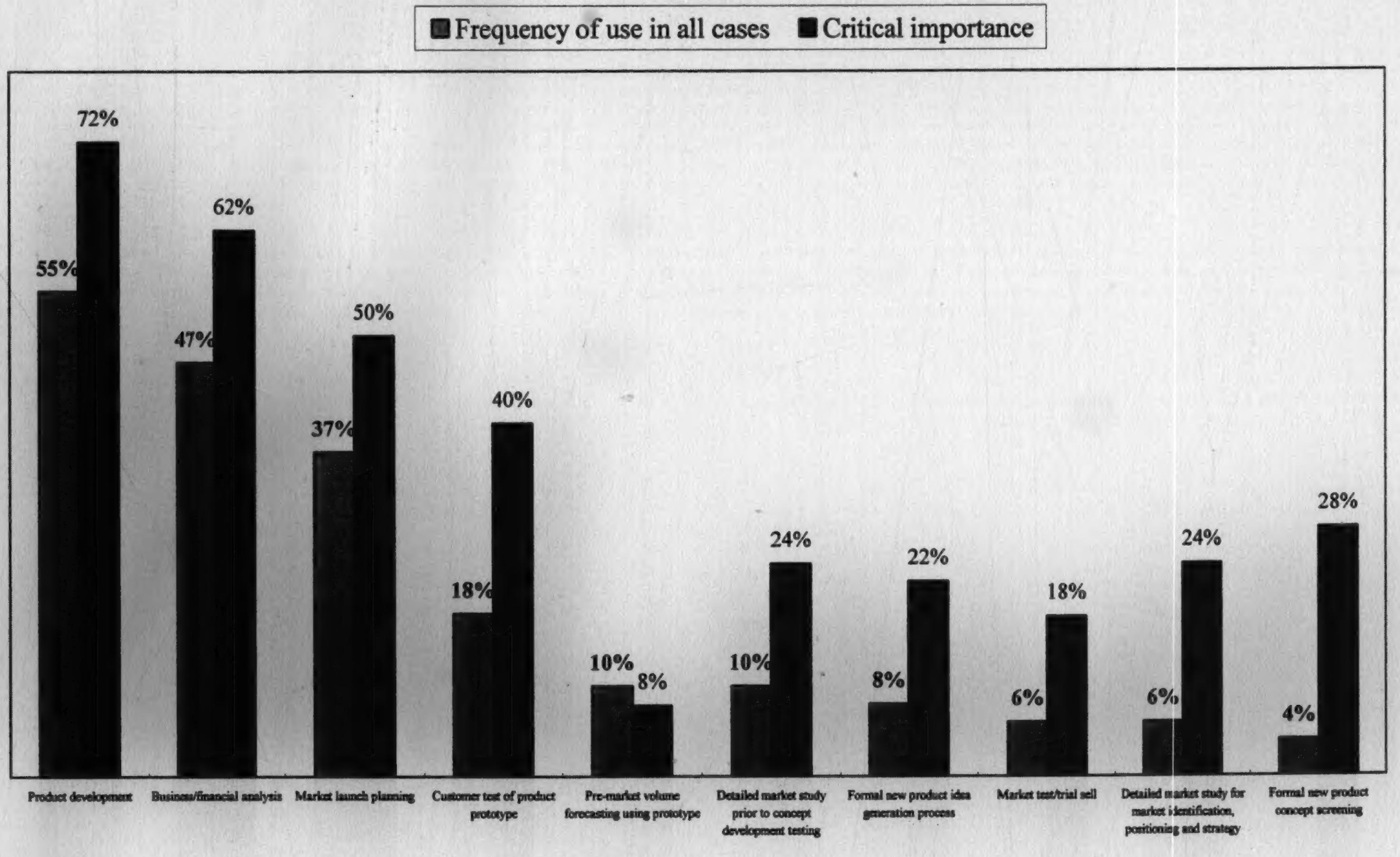
Standard deviations are given in parentheses.

Table 3-3
Frequency of Use and Importance of NPD Activities

Activities	Frequency			Importance			
	All cases	Some cases	Very few cases	Critical	Important	Marginal	Not at all
	%	%	%	%	%	%	%
Formal new product idea generation process	8	35	53	22	52	22	4
	(30)	(49)	(20)	(19)	(50)	(26)	(3)
Formal new product concept screeng	4	49	33	28	58	10	4
	(43)	(40)	(15)	(40)	(38)	(13)	(6)
Detailed market study prior to concept development testing	10	29	51	24	40	34	2
	(4)	(53)	(41)	(14)	(43)	(31)	(6)
Detailed market study for market identification, positioning and strategy	6	41	45	24	56	18	2
	(31)	(41)	(23)	(34)	(43)	(11)	(8)
Business /financial analysis	47	29	24	62	30	6	2
	(70)	(23)	(6)	(55)	(38)	(3)	(4)
Product development	55	33	8	72	26	0	2
	(79)	(17)	(3)	(71)	(23)	(1)	(1)
Customer test of product prototype	18	33	43	40	48	10	2
	(52)	(27)	(18)	(48)	(37)	(9)	(1)
Pre-market volume forecasting using prototype	10	20	41	8	44	44	4
	(20)	(33)	(47)	(9)	(47)	(27)	(16)
Market test/trial sell	6	16	51	18	26	46	10
	(22)	(31)	(42)	(20)	(42)	(18)	(16)
Market launch planning	37	29	31	50	40	6	4
	(54)	(30)	(14)	(41)	(45)	(12)	(1)

The corresponding results of Mahajan and Wind (1992) are given in parentheses.

Figure 3-2
Relationship between Frequency of use and Critical Importance



The models and/or methods used and the degree of satisfaction

Respondents were asked which of 13 specific NPD models and methods were used in their organization during the previous three years. These consisted of the 11 alternatives investigated by M&W, as well as stated choice/preference models and ethnographic/observation usage tests, which have attracted more attention in recent years. Table 3-4 lists these models and methods in order of the percentage of companies who report using them. The NPD method used most widely in the sample is concept tests, reported by 77% of the respondents. Focus groups are the second most widely used method, reported by 66%. Limited rollout is the other method used by more than half of the companies. Perhaps surprisingly, more companies now use stated choice/preference models (28%) than use traditional conjoint analysis (19%). Besides concept tests, the methods included in M&W that have become much more popular are product life-cycle models (38% cf. 8%), quality functional deployment (26% cf. 9%) and attitude and usage studies (36% cf. 19%).

Table 3-4
Models & Methods Frequently Used in NPD Process

Model/method	% of companies	% in Mahajan & Wind
Concept tests	77	26
Focus groups	66	68
Limited rollout	53	42
Product life-cycle models	38	8
Show tests and clinics	36	22
Attitude and usage studies	36	19
Stated choice/preference models	28	NA
Quality function development (QFD)	26	9
Ethnographic/observation usage research	23	NA
Traditional conjoint analysis	19	15
Home usage test	19	9
Delphi	6	9
Advanced product quality planning	2	NA
Synectics	0	8

Satisfaction with the models and methods that have been used was measured using a five-point Likert scale with categories labeled from completely satisfied to completely dissatisfied. As shown in Table 3-5, completely satisfied is not reported that frequently. However, users are generally more satisfied with concept tests, focus groups, limited rollout, and attitude and usage studies. They are less satisfied with product life-cycle models, stated choice/preference models, and the less commonly used traditional conjoint analysis, home usage tests, and Delphi. Thus, there is some relationship between frequency of use and relative satisfaction with a new product development model or method.

(Insert Table 3-5 about here)

Current Status of Concept Testing

As shown in Table 3-6, the number of concept tests conducted per year ranged widely by company type from 265 for package goods, 37 for others, 31 for durables, 9 for services, and 7 for industrial. However, only a small number of these concept tests were of innovative products. Moreover, as shown in Table 3-7, the average number of new products introduced is only 19% (13/67) of the average number of concepts tested. The "weed out" rate is especially high for package goods, where the number of new products that are introduced is only 7% of the number of products that are tested. In contrast, concept tests are apparently not used to screen out weaker ideas for new durables and industrial products.

(Insert Table 3-6 about here)

(Insert Table 3-7 about here)

Most Recent Concept Testing Project

To obtain a more in depth look at concept testing design, the respondents were asked to provide more detailed information on the organization's most recent traditional or conjoint concept testing project where the outcome had been determined and a resulting management decision had been made. The specific concepts tested in the most recent concept testing projects were 22% industrial, 22% service, 26% package goods, 22% durables and 9% others, as shown in Table 3-9. About 65% of these most recent projects tested incrementally new product concepts (incrementally new to the market, improvement to an existing product, repositioning of an existing product or extension of an existing product). The remaining 35% of the projects tested concepts that are radically new to the market. Because of the long running concern about the applicability of concept testing for radical innovations, subsequent analyses compare the tests of incremental and radical concepts.

Table 3-5
Level of Satisfaction with the Various Models & Methods

Model/method	No. of users	Level of satisfaction (% of users)				
		Completely satisfied	Somewhat satisfied	Neutral	Somewhat dissatisfied	Completely dissatisfied
Concept tests	37	22	70	3	5	0
Focus groups	31	16	71	3	6	3
Limited rollout	25	24	60	12	4	0
Product life-cycle models	18	6	39	39	17	0
Show tests and clinics	17	6	82	12	0	0
Attitude and usage studies	17	24	59	18	0	0
Stated choice/preference models	13	8	46	46	0	0
Quality function development (QFD)	12	25	42	25	8	0
Ethnographic/observation usage research	11	18	64	18	0	0
Traditional conjoint analysis	9	0	67	33	0	0
Home usage test	9	33	22	44	0	0
Delphi	3	0	33	67	0	0
Advanced product quality planning	1	0	0	100	0	0

Table 3-6
Annual Average Number of Concept Tests

Company type	2002		2003		2004		Annual Average	
	All concept tests	Tests for innovative products	All concept tests	Tests for innovative products	All concept tests	Tests for innovative products	All concept tests	Tests for innovative products
Package	380	1.8	336	2.2	154	1.5	265	1.8
goods	(747)	(2.9)	(815)	(1.0)	(423)	(1.6)		
Durables	36	4.0	38	4.5	24	3.7	31	4.0
	(57)	(4.5)	(55)	(3.7)	(38)	(3.4)		
Industrial	6	2.4	7	2.7	7	3.3	7	2.8
	(6.9)	(3.5)	(6.4)	(3.2)	(6)	(3.1)		
Service	7	2.8	9	2.8	12	3.4	9	3.0
	(5)	(1.7)	(7.7)	(2.5)	(10)	(2.6)		
Others	26	5.6	33	8.9	51	11.8	37	9.0
	(62)	(13.9)	(85)	(21.2)	(113)	(19.0)		
Total	69	3.7	78	5.0	55	5.7	67	4.9
	(278)	(8.4)	(344)	(12.8)	(202)	(11.4)		

Standard deviations are given in parentheses.

Table 3-7

Comparison of Annual New Product Introductions and Concept Tests

Company type	2002		2003		2004		Annual Average	
	New products	Concept tests	New products	Concept tests	New products	Concept tests	New products	Concept tests
Package goods	28	380 (7%)	17	336 (5%)	14	154 (9%)	19	265 (7%)
Durables	41	36 (-)	24	38 (63%)	27	24 (-)	29	31 (94%)
Industrial	12	6 (-)	12	7 (-)	13	7 (-)	12	7 (-)
Service	5	7 (71%)	4	9 (44%)	5	12 (42%)	5	9 (56%)
Others	5	26 (19%)	6	33 (18%)	8	51 (16%)	7	37 (19%)
Total	14	69 (20%)	12	78 (15%)	13	55 (24%)	13	67 (19%)

Success rates (=1 minus weed out rate) are given in parentheses.

Analysis Method - Statistical Tests for the Comparison

I have 33 tests of incremental concepts and 18 tests of radical concepts. The two groups are independent then the independent sample *t*-test (Snedecor and Cochran 1989) can be used to determine if two population means (incremental versus radical) are equal. The independent sample *t*-test is defined as:

H$_0$: $\mu_1 = \mu_2$

Ha: $\mu_1 \neq \mu_2$

Test statistic: $T = \dfrac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_1^2/N_1 + S_2^2/N_2}}$ (McCabe and Moore 2005)

where μ_1 and μ_2 are the population means for incremental concepts and radical concepts, N_1 and N_2 are the sample sizes, $\overline{Y}_1$ and $\overline{Y}_2$ are the sample means, and S_1^2 and S_2^2 are the sample variance. Approximately normal distribution of the measure in the two groups is assumed. The independent sample *t* test also assumes homogeneity of variance in the two groups. Homogeneity of variances is tested by "Levene's Test for Equality of Variances", with F value and corresponding significance. These are part of SPSS output for two independent sample t-tests. The t-test may be unreliable when the two samples are unequal in size and also have unequal variances (Gardner 1975). I used the conventional level of 0.05 as the significance level.

If the data are proportions rather than means, the comparison for the difference between population proportions is just as I did for population means because a proportion is simply the mean of a dichotomized variable, measure 0 for one category, and 1 for the other. The mean for this variable is simply the proportion of those in category 1 (or percent, if multiplied by 100). In the special case of proportions, the standard error of a sampling distribution can use simplications in computations. The variance of a dichotomous variable when *P* is the proportion in one category and (*1-P*) is the proportion in the other is simply, *P (1-P)*, so the standard deviation becomes square root of *P (1-P)* divided by sample size (*N*). The test statistic becomes:

Test statistic: $Z = \dfrac{\overline{P}_1 - \overline{P}_2}{\sqrt{(\overline{P}_1(1-\overline{P}_1))/N_1 + (\overline{P}_2(1-\overline{P}_2))/N_2}}$ (McCabe and Moore 2005) where $\overline{P}_1$ and $\overline{P}_2$ are percentages for incremental concepts and radical concepts.

It is important to note that here I run a greater risk of the small sample being unusual just by chance according to the central limit theorem. Because when the sample size is small, it may not be unreasonable that I could draw too extremely rare observations that are far from the population mean, and thus give a sample mean that is far from the true mean.

Two independent sample *t*-tests are available in just about all general-purpose statistical software programs. In my analysis, I used SPSS two sample *t*-test and the alternative is an Internet statistical tool SISA (Simple Interactive Statistical Analysis) serviced by quantitative skills to conduct two independent sample *t*-tests.

1. The Objectives of the Concept Testing Project

Table 3-8 reports the percent of respondents indicating particular objectives were important in their most recent concept-testing project. The most commonly reported important objective was to develop further the original idea (cited by 81% of the respondents). Other commonly reported important objectives were to estimate the concept's market potential (70%), eliminate poor concepts (66%), identify the value of concept features (66%) and help identify the highest potential customer segment (53%). Clearly, concept tests usually have multiple objectives, some of which imply the objects to be scaled in a concept test are not just concepts. When concept tests are used to estimate the concept's market potential or to eliminate poor concepts, the concepts themselves are the object of measurement. But, when concept tests are used to help identify the highest potential customer segment, they are being used to identify how different customer segments respond to particular concepts, which means that the interaction between concepts and segments is really an object of measurement. In other cases, when concept tests are used to develop further the original idea, identify the value of concept features and provide diagnostic information, practitioners need to identify the value of different aspects (e.g., attributes and features) of the tested concepts, the interaction between concepts and aspects is the object of measurement. Thus these results indicate that not only concepts, but also the interactions between concepts and aspects and between concepts and segments are important objects of measurement in concept testing. The objectives of tests of incremental and radical concepts seem similar.

However, looking at the last column of Table 3-8, which identifies the implied objects of measurement for the different measurement objectives, objectives which imply interactions are the objects of measurement are a little more prevalent in tests of radical concepts, but not statistically significant at 0.05 or 0.1 level.

Table 3-8
Objectives of the Concept Testing Project

Objectives	Percent respondents			Implied object of measurement
	Total	Incremental	Radical	
Develop further the original idea	81	79	88	Concepts by aspects
Estimate the concept's market potential	70	68	81	Concepts
Eliminate the poor concept(s)	66	68	63	Concepts
Identify the value of concept features	66	61	81	Concepts by aspects
Help identify the highest potential customer segment	53	54	63	Concepts by segments
Generate an estimate of the sales or trial rate	32	32	25	Concepts
Provide diagnostic information	28	32	25	Concepts by aspects

Concept tests need to provide adequate data quality for each of the objects of measurement. A data collection design that successfully scales concepts does not guarantee similar success when scaling other objects of measurement (e.g., scaling concepts by segments).

2. The Design of the Concept Testing Project

Table 3-9 breaks down the characteristics of the concept test design by the newness of the tested concepts. Note that durables make up a larger share of the incremental concept tests while services make up a larger share of the radical concept tests.

The number of concepts evaluated per project ranged from one to 30, with a mean of six. Forty-six percent of the projects were monadic tests while the rest were comparative tests. Seventy percent of the concept testing projects included pricing information, 27% didn't include pricing information and 2% of the projects included concepts with and without pricing information. There were no significant differences between tests of incremental and radical concepts in these three areas.

Concepts can be presented in several ways. This survey found concepts were most commonly presented as stripped descriptions (38%) or stripped with visual representation (31%). Stripped descriptions were even more often used for radical concepts (50%) and stripped with visual representation were more often used for incremental concepts (37%), where others (15%), likely prototypes, were used more often.

Table 3-9 also reports that the criteria most often used to select respondents for the concept tests were product class usage (used in 57% of tests) and specific product usage (43%). For tests of radical concepts, there was more use of the lead user criteria (56%), as expected from Supposition 1a. Perhaps surprisingly, there was no greater use of an innovativeness or influential/market maven criteria for radical concepts.

The number of respondents used to evaluate each concept averaged 92 with a standard deviation of 170. Consistent with Supposition 1b, significantly larger numbers of respondents evaluated each concept in tests of radical concepts. Whereas an average of only 39 respondents evaluated each incremental concept, an average of 195 respondents evaluated each radical concept. Moreover, the sample size for radical concepts was far more varied (a standard deviation of 250 compared with 66), suggesting there is still no recognized sample size norm for tests of radical concepts.

(Insert Table 3-9 about here)

Table 3-10 reports on the specific unstructured and structured questions used to assess respondents' reaction to concepts. Liking (used in 77% of tests), disliking and purchase intent are the most popular unstructured or open-ended questions. The most popular structured rating scale questions are purchase intent (used in 60% of the tests), comparison with current offering (49%), liking (44%), comparison with expectation (42%) and uniqueness (35%). Consistent with Supposition 1c, there is more use of both unstructured and structured measures for concept testing of radically new products than for concept testing of incrementally new products. For example, there is significantly greater use of an unstructured liking question, which becomes virtually universal, and there is significantly more use of structured purchase intent and disliking questions.

(Insert Table 3-10 about here)

Table 3-9
The Design of Tests for Incremental and Radical Concepts

Design characteristics	Total	Incremental	Radical
Type of products			
- Industrial (%)	22	21	25
- Service (%)	22	18	31
- Package goods (%)	26	21	25
- Durables (%)	22	29	13
- Others (%)	9	11	6
How many concepts evaluated in a single project			
Average (n)	6	6	6
Monadic test or comparative test			
Monadic test (%)	46	46	47
Comparative test (%)	54	54	53
Pricing information			
Pricing information included (%)	70	67	73
Pricing information not included (%)	27	30	27
Both (%)	2	4	0
Form of concept presentation			
Stripped description (%)	38	30	50
Embellished description (%)	9	7	13
Stripped with visual representation (%)	31	37	19
Rough mock advertisement (%)	4	4	6
Fully finished advertisement (%)	9	7	13
Others (%)	9	15	0
Respondent selection criteria			
Product class usage (%)	57	50	9
Specific product usage (%)	43	46	38
Innovativeness (%)	14	15	13
Lead user criteria (%)	36	23	56
Influentials/market maven criteria (%)	26	27	25
Demographics (%)	26	31	19
Lifestyle group membership (%)	10	12	6
General population (%)	5	4	6
How many respondents			
Average (n)	92	39	195*
Standard Deviation	170	66	250*

*$p < 0.05$.

Table 3-10
Response Scales Used

Questions	Total	Incremental	Radical
Unstructured/qualitative			
Purchase intent (%)	63	56	75
Comparison with current offering (%)	53	59	44
Liking (%)	77	67	94*
Comparison with expectation (%)	56	59	50
Uniqueness (%)	44	37	56
Disliking (%)	63	56	75
Believability (%)	35	30	44
Problem solving ability (%)	30	33	25
Others (%)	7	11	0
How many unstructured questions			
Average	4.3	4.1	4.6
Structured/quantitative			
Purchase intent (%)	60	48	81*
Comparison with current offering (%)	49	41	63
Liking (%)	44	37	56
Comparison with expectation (%)	42	33	56
Uniqueness (%)	35	33	38
Disliking (%)	26	15	44*
Believability (%)	26	26	25
Problem solving ability (%)	23	22	25
Others (%)	12	15	6
How many structured measures			
Average	3.2	2.7	3.9
Response scale format			
Numerical categories only (%)	22	36*	0
Categories with end-point labeled only (%)	13	14	8
Categories with end & mid-point labeled (%)	6	0	17*
Categories with all points labeled (%)	22	14	33
Both numerical categories & verbal labels (%)	25	18	42
Continuous response scale (%)	3	5	0
Other (%)	9	14	0

*$p<0.05$.

For structured questions, five-point or ten-point scales are used in most projects. As shown in Table 3-10, the most popular response scale formats use both numerical categories and verbal labels (25%), numerical categories only (22%) and categories with all points labeled (22%). Tests of radical concepts do not use scales with numerical categories only (0%), and are significantly more likely than tests of incremental concepts to use scales combining both numerical categories and verbal labels (42%) and scales with the end-point and the mid-point labeled (17%). Tests of incremental concepts most often use numerical categories only (36%). The continuous response scale is rarely used for either type of concept test.

3. Data Collection Methods

Concept-testing data can be collected in several different ways. As summarized in Table 3-11, the most popular collection methods are focus groups (33%), in-home interviews (23%) and mall intercepts/central location (16%). No one reported they used a mail survey for their concept-testing project. These results indicate that face-to-face data collection methods still predominate. There is no significant difference in the data collection methods used for testing the two types of concepts.

Table 3-11
Data Collection Methods

Data collection method	Total	Incremental	Radical
Focus groups (%)	33	33	31
In-home interviews (%)	23	19	31
Mall intercepts/Central location (%)	16	19	13
Telephone interviews (%)	14	19	6
Mail survey (%)	0	0	0
On-line survey (%)	14	11	19

It is interesting to note that on-line concept testing was used for 14% of the tests. However, a follow-up question revealed that 40% of the companies have conducted an on-line concept-testing project in the last three years, and on average 16% of concept testing projects were conducted on-line during that period. The three primary reasons provided for the use of on-line concept testing were (1) cost-effectiveness, (2) takes less time and (3) ease of use.

4. Methods for Analyzing the Data

Table 3-12 displays the methods used to aggregate the responses obtained and to judge the outcome of the concept test for the quantitative questions. The percent Top 2 Box Scores and the rating scale mean are the most commonly employed methods used to summarize the ratings scale data. The traditional but simplistic percent Top 2 Box Scores is used less and the more appropriate mean and median both are used more for tests of radical concepts than for tests of incremental concepts. While the directionality of these three differences is consistent with Supposition 1d, none is significant at conventional levels. The comparison standards used in interpreting these summary measures are most often company norms, particularly for radical concepts. Overall it appears practitioners primarily rely on simple measurement heuristics to assess the results of their concept tests.

Table 3-12
Comparisons of Methods for Analyzing the Data

Assessment method	Total	Incremental	Radical
Method of aggregation			
Percent top-box score (%)	24	25	23
Percent top-2-box scores (%)	45	50	38
Rating scale mean (%)	36	30	46
Rating scale median (%)	21	15	31
Comparison standard employed			
Comparison with company norms (%)	39	30	54
Comparison with industry norms (%)	12	10	15
Comparison with research supplier norms (%)	12	15	8
Other (%)	3	5	0

5. The Level of Satisfaction with the Predictive Performance

Satisfaction with the predictive performance of the organization's current approach to concept testing was below that previously reported for satisfaction with concept tests (as shown in Table 3-5, 22% of the respondents reported completely satisfied, 70% somewhat satisfied, 3% neutral, 5% somewhat dissatisfied and 0% completely dissatisfied). As shown in Table 3-13, only 7% of respondents were completely satisfied with the predictive performance of their concept tests, while 57 % were somewhat satisfied, 17% were neutral, and 19% were somewhat dissatisfied. It looks like respondents were no more satisfied with the predictive performance of their tests of incremental concepts than of radical concepts.

Table 3-13
Level of Satisfaction with Predictive Performance

The level of satisfaction	Total	Incremental	Radical
Completely satisfied (%)	7	4	14
Somewhat satisfied (%)	57	54	64
Neutral (%)	17	21	7
Somewhat dissatisfied (%)	19	21	14
Completely dissatisfied (%)	0	0	0

Looking at Supposition 2 (New product managers who have tested innovative concepts would be less satisfied with the predictive performance of their concept testing than managers who have tested incremental new products.) again, it would only seem to be a reasonable supposition if expressed in terms of controlling for differences in methods used for incremental and radical concepts. Then, instead of Table 3-13, I ran a linear regression model with satisfaction as the dependent variable and newness of concepts (NEWNESS) and other characteristics, such as sample size (NPERS), the number of response questions (NOPEN and NCLOSE), as the independent variables. All requested variables entered the full model in which I attempted to see if there is an effect of newness of concepts. Table 3-14 summarized the coefficient estimates. It appeared that except for the intercept, all independent variables are not significantly different from zero. In addition, the R squared is .112, suggesting that the model only accounted for 11.2% of the total variance. The model F statistic is 1.165 (p value is .342), indicating that the linear relationship between the design variables and satisfaction is not significant. Though the regression model is not that successful, the results revealed that the newness effect is not significant after controlling for differences in the rigorousness of the testing procedures, such as sample size and the number of structured response measures used. Note that other curve estimation regression models were also investigated, but none would fit better than the linear model.

Table 3-14
Summary of the Results from the Regression Analysis

	Unstandardized Coefficients	Std. Error	Standardized Coefficients	*t*	Sig.
(Constant)	3.854	.393		9.799	.000
NPERS	-.000	.001	-.038	-.218	.829
NEWNESS	.176	.157	.198	1.119	.271
NOPEN	-.113	.080	-.239	-1.416	.165
NCLOSE	-.087	.062	.239	1.407	.168

Table 3-15
Level of satisfaction with various concept testing designs

	Reported level of satisfaction				
Design characteristics	Completely Satisfied	Somewhat Satisfied	Neutral	Somewhat Dissatisfied	Completely Dissatisfied
How many structured measures					
Average	4	4	2	3	-
How many respondents					
Average	101	114	17	98	-

Table 3-15 reports the results from an exploration of possible relationships between the level of satisfaction and aspects of concept testing design. Respondents who were more satisfied appear to have used more structured responses and used more respondents to evaluate each concept.

In their response to the unstructured open-ended questions, respondents identified several shortcomings they could see with concept testing. The major problem they mentioned is the variable predictive validity of concept testing. This raises the question of whether practitioners adequately track their results and compare results over time and against subsequent projects. Two other shortcomings frequently mentioned by respondents were the applicability of concept testing to “new to the world” products and respondent selection bias.

The respondents also provided some suggestions to improve concept testing. Two common themes appeared in their answers, namely (1) to make more use of on-line testing, and (2) to make more use of virtual testing, such as measuring preferences in the context of competitive sets and in more natural shopping environments. Other specific suggestions made to improve concept-testing design and methods were to include more careful screening of participants, larger samples, better financial models providing probabilistic estimates of expected outcomes, and oblique testing for radically new products.

3.2 The Study of Marketing Research Firms

I supplemented the survey of new product managers by contacting marketing research firms, who provide concept-testing services and asking them for any publicly available evidence as to the reliability and validity they claim for their concept testing services.

Study Approach

To begin, 100 research consultants were identified from three sources, namely, relevant categories of the Product Development & Management Association (PDMA) website, a Google search of “new product concept testing” and the relevant categories of Marketing Services Directory (AMA M Guide).

First, I coded (documented) what they say publicly about the concept testing service they offer. Second, I sent them a request for information about their concept testing services they would provide to a potential client. This request asked them to confirm they offer product concept testing services, what types of tests they provide (quantitative or qualitative, on-line or traditional), a general question about what they would recommend when setting up a concept test for a particular product, and for any information they provide about the reliability and validity track record of their concept testing services.

Evidence for Reliability and Validity

Responses were obtained from 31 of the 100 consultants. Among those responding, 17 consultants clearly indicated that they have no record of the reliability or validity of their concept testing work and rely on their clients to provide that type of information for themselves. The remaining 14 consultants claimed to have reliability and validity data for their concept testing applications, but only seven had evidence for their reliability and validity track record and were prepared to share details about the nature of the data, as detailed below.

Case studies – Two consultants accumulated a database of quantitative case studies where they accurately forecasted sales of new technologies, some over periods as long as 5 years.

Normative benchmarks – One consultant claimed superior performance for products that test well and pass a hurdle rate based on an internal benchmark or industry normative benchmarks and go on to be developed and launched. Typically those products using normative benchmarks for decision making have a 20-50% higher success rate (measured in product lifetime revenues) than untested products.

Quantitative models – The other four consultants use predictive mathematical models to calculate an overall "SuccessScore" for each concept. This score is then compared to the action standards to determine if the new product concept warrants further development. Or they use the data they collect in volumetric forecasting models that have been validated to predict sales.

Reasons for the Lack of Reliability and Validity Evidence

Other consultants provided the following explanations for their inability to provide evidence for reliability and validity:

1. Such information is confidential.
2. Clients rarely share it with their suppliers. Since clients are often unwilling to share information on the market success or failure of their new products that were

concept tested, the consultant's ability to validate results is limited.

3. Clients do not often test the same concept twice, so little reliability data is available.

3.3 Conclusions on the State of Contemporary Practice

Given the proliferation of new product introductions and intensifying domestic and global competitiveness in the new product market, the survey serves to shed light on the state of contemporary practice related to the use of various models and methods of concept testing, their problems and desired improvements, and the evidence for reliability and validity. The sample responses lead to the following major conclusions:

First, the characteristics of a typical concept test project can be summarized as follows: (1) has multiple objectives, but the primary objective is to develop further the original idea; (2) tests incrementally new product concepts; (3) is equally likely to be a monadic or comparative test; (4) includes pricing information in the tested concepts; (5) presents the concept stimuli as stripped descriptions; (6) uses about 92 respondents for each concept; (7) uses both unstructured and structured measures; (8) employs scales with both numerical categories and verbal labels; (9) selects respondents by product class usage, specific product usage or lead user criteria; (10) assesses the outcome using percent Top 2 Box Scores and/or rating scale mean compared with a company specific norm; and (11) collects data by face-to-face interview.

Second, concept testing was expected to be more rigorous for radical new products. Thus testing of radical new products was expected to more carefully screen respondents (Supposition 1a), employ larger samples (Supposition 1b), collect more detailed information (Supposition 1c), and employ more sophisticated analysis methods (Supposition 1d). I found statistically significant evidence to support most of these suppositions. For radical concept tests, lead user criteria are more commonly used. An average of only 39 respondents per concept was reported in incrementally new product concept tests compared with 195 for radically new product concept tests. There is more use of both unstructured liking and structured purchase intent and disliking measures for radical concept tests. There is more use of company norms to judge the outcome for radical concept tests, but otherwise the analysis methods are not more sophisticated.

Third, and contrary to Supposition 2, respondents are no more satisfied with the predictive performance of their current concept testing approach when testing incremental concepts than when testing radical concepts. One explanation could be that managers who are testing radical concepts employ far larger samples of respondents, and so are more confident about the results. Alternatively, they could simply have lower expectations for their predictive performance. Other exploratory findings suggest respondents are more satisfied with concept tests when using more than four structured questions and when using more respondents (greater than 100) to evaluate each concept.

I summarize whether the suppositions I made are consistent with the survey findings. The findings supported the supposition 1, including supposition 1a, 1b, 1c and 1d, that concept testing of radical product innovations is more rigorous than that of incrementally new products. Not supported is supposition 2 that new product managers who have tested innovative concepts are less satisfied with the predictive performance of their concept testing than managers who have tested incremental new products.

Finally, it is clear that most practitioners prefer to keep their concept testing information proprietary. Almost all evidence for reliability and validity is kept confidential, so there is little public information on concept test performance. Thus surprisingly little seems to have been learned about how exactly concept tests should be designed, despite the thousands of concepts that are tested every year. This makes primary academic research into the measurement issues in concept testing more necessary and valuable.

3.4 Limitations

The survey of new product managers provides a descriptive study of how companies currently conduct product concept testing. Since limited resources meant the study is based on a relatively small sample of new product managers, it could certainly be argued that the survey results may not fully represent current users of concept testing across all types of industries and firms. However, the respondents obtained by surveying members of a NPD related association would seem likely to be as close to the current state of the art as any other sample of users.

In addition, the results are based on only 51 respondents and some respondents failed to reply to some of the survey questions for confidentiality reasons. Many of the subcategories that are of interest in the analysis contain only a few observations. For example, the tests of the expected differences between practices for incremental and radical new concepts could often only detect particularly strong effects. Finally, the study of research consultants is exploratory and qualitative in nature.

In summary, the survey of new product managers and research consultants is all about describing the current state of concept testing practice. Limited resources and lack of a solid theoretical framework meant it could not fully answer all questions and no prescriptive conclusions could be drawn for managers. However, the findings of the study do help us understand the current state of concept testing practice. Moreover, it illustrates how necessary and valuable academic research addressing the measurement issues in concept testing will be. In the following chapters, secondary and primary studies are used to answer the questions of what distinguishes the better tests and how those relate to consumer trial – both of which are central for practice.

Chapter 4 Conceptual Framework

This chapter first introduces the generalizability theory (G theory) approach to measurement. Then, it presents a conceptual framework for the factors influencing testing responses from a G theory perspective, and it identifies the implicit assumptions currently being made about these factors in the conduct of concept testing.

4.1 Generalizability Theory Approach

G theory, pioneered by Cronbach and his colleagues (Cronbach, et al. 1972), has long been identified as a superior approach to measurement issues in marketing (e.g., Peter 1979; Rentz 1987; Finn and Kayande 1997). G theory uses analysis of variance to provide estimates of the observed score variation due to multiple sources (e.g., variation due to concepts or due to the use of different respondents, items, occasions). It explicitly recognizes the fact that measurement takes place for multiple objects of measurement and over multiple facets of generalization, reflecting the multi-faceted nature of measurement error. The generalizability of concept testing depends on the size of errors that are relevant for the particular object of measurement. G-theory provides object specific G-coefficient criteria for assessing the psychometric quality of research data.

A Review of Basic Concepts in G Theory

G theory is the most comprehensive approach to assessing the reliability of measurement. It was originally developed for educational testing by Cronbach and his colleagues (1972) and was recently updated by Brennan (2001a). Peter (1979) had the foresight to identify G theory as being of potential interest to marketing scholars. Rentz (1987, 1988) provided a full introduction and presented some demonstration results for the generalizability of some consumer scales using data provided by students. Finn and Kayande (1997) noted that G theory combines a powerful conceptual framework and a set of empirical procedures, which gives the researchers the ability to separate different sources of variation and to estimate the magnitude of the variance components using analysis of variance.

The basic philosophy underlying G theory is that "an investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs" (Cronbach, Rajaratnam and Gleser 1963, p.144). Because some of the terminology in this research is unique to G theory, I first introduce some basic definitions of terms.

Object of measurement An object of measurement is a factor (e.g., concepts/products in concept testing, firms, advertisements, brands) whose levels must be scaled by the measurement instrument.

Facet and conditions of a facet In G theory terminology, a ***facet*** is a set of conditions of measurement of the same kind. For example, the set of relevant measurement occasions might constitute a facet in a particular study. A facet is analogous to a factor in analysis of variance. Conditions of a facet are analogous to levels of a factor in analysis of variance.

Population In G theory the word *universe* is reserved for conditions of measurement, while the word *population* is used for the objects of measurement. If the object of measurement in a concept test is concepts, managers need to specify the population of concepts in which they would like to generalize the test results.

Universe of admissible observations The universe of admissible observations is all possible observations that a test user would consider acceptable substitutes for the observation in hand. G theory views a behavioral measurement as a sample from a universe of admissible observations. The universe is characterized by one or more sources of error variation or facets. For example, assuming the object of measurement is concepts, the facets in a typical concept test (when 100 respondents evaluate the 10 tested concepts using 6 evaluation items) are respondents and items. The universe for such a concept test consists of all combinations of the levels of respondents and items. The concept testing results represent a sample from the universe of admissible observations. The decision maker intends to generalize the testing results to the entire universe of admissible observations.

Universe of generalization The universe of generalization is the set of all such conditions of measurement over which the investigator wishes to generalize. For example, marketing researchers seldom are interested in generalizing a measurement over only the particular occasion on which the measurement is taken (i.e., 1:03 p.m. on June 1). Interest is usually in generalizing to the set of all such occasions (usually within some time interval). There is a universe of occasions to which one wishes to generalize. Similarly, one might wish to generalize over a universe of items, interviewers, and situations of observation.

A facet of generalization and a facet of differentiation *A facet of generalization* is a facet over which one wishes to generalize. In other words, it is a set of conditions that contribute unwanted variation (measurement error) to observations in a study. Therefore, the measurement instrument should minimize variance arising from these sources. Facets of generalization contribute random error but also may contribute systematic error. *A facet of differentiation* is a set of objects that are to be compared in a study. This is the facet that contributes desirable variance. The purpose of a study is to distinguish between levels of this facet and therefore the measurement instrument should maximize variability arising from a facet of differentiation. For a particular scale, the facets of differentiation and generalization might differ depending on the purpose of measurement.

G study and D study *A Generalizability study* (G study) is designed to investigate the various sources of measurement error that arise from various conditions of measurement. A *Decision study* (D study) collects data for a particular decision. The findings of a G study can be used to help in the design of a D study. Generally, however, the G study is more comprehensive. The conditions of measurement are varied systematically so that the contribution of each facet (the facets of generalization and differentiation) is estimated. This is done by examining the variance components associated with each of the facets. This information can be used in instrument refinement or in the design of subsequent D studies. The D study may contain fewer conditions, depending on the purpose of the study and the results of the G study. For example, if the G study shows that certain facets contribute little error, the number of conditions of those facets can be reduced in subsequent D studies with little loss of generalizability. Resources would be better spent to increase the sample of conditions contributing larger amounts of error so that generalizability is increased. The ability to predict and control the sources and magnitude of measurement error in subsequent D studies is unique to G theory and should be of great practical importance to marketing researchers.

Relative error and absolute error G theory distinguishes two types of error. The type of error used depends on the type of interpretation of the scores the analyst will make. Relative error variance is appropriate when a relative interpretation is made of the scores. For example, five concepts might be ranked according to their scores on a concept test and two highest ranked concepts considered for further development. This decision is a relative one and the relative error is appropriate. Most marketing decisions are relative decisions, so the relative error is appropriate in most cases. Absolute error variance is appropriate when the scores are to be interpreted in an absolute sense. For example, a concept may be considered for further development only if its score on a concept test exceeds some threshold score. The decision is an absolute one and the absolute error is appropriate. The absolute error indicates how far measures are likely to depart from their "true" values.

Generalizability coefficient $E\rho^2$ A Generalizability coefficient (G coefficient) is the ratio of universe score variance to itself plus relative error variance:

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)}$$

where $\sigma^2(\tau)$ is universe score variance and $\sigma^2(\delta)$ is relative error variance.

The G coefficient is the analogue of a reliability coefficient in classical test theory. The required value for this coefficient determines the amount of data that need to be collected in a subsequent decision study.

The design of a G study is analogous to the design of experiments. Facets are analogous to factors in analysis of variance and the conditions of a facet are analogous to levels of a factor. The conditions of a facet may be either fixed or random. Facets may be crossed or nested. The investigator obtains measurements under various conditions of all relevant facets of generalization and differentiation. A G coefficient, which characterizes the psychometric quality of measurement, may be computed for each facet or combination of facets of generalization. The coefficient indicates the extent to which the observations can be generalized to the universe of similar observations.

Procedures of G theory approach In the G theory approach, an initial G study is conducted in which data are collected to determine how sensitive the construct being measured is to the levels of different facets in the measurement environment. Then, when subsequent managerial decisions need to be made, knowledge of the extent of variation across and within facets can be used to determine how many observations will be necessary to draw managerial conclusions with a required degree of reliability in a D study. If cost information is also available, it is possible to identify the most cost effective designs for each subsequent decision study. Thus, the G-theory approach is most beneficial in a programmatic research context, rather than for one-off research projects.

Applying G Theory to Concept Testing

When using G theory, concept test scores are interpreted in light of the estimated variance components for all the sources that can contribute to the observed variation. Consider a concept test in which respondents complete response tasks for concepts in various situations. From a G theory perspective, the response variation is attributed to the concepts, respondents, tasks and situations that constitute facets of variation in the evaluation study design.

Presumably, one would accept as admissible the response of any respondent (in the population) to any concept (in the concept universe) on any item (in the item universe) in any situation (in the situation universe). If so, the population and universe of admissible observations are crossed. G-theory assumes a random effects model such that the observed response score X_{crsi} provided by a respondent, *r*, responding to concept, *c*, in situation, *s*, for response task (item), *i*, can be expressed as

(1)

$$X_{crsi} = \mu + \nu_c + \nu_r + \nu_s + \nu_i + \nu_{cr} + \nu_{cs} + \nu_{ci} + \nu_{rs} + \nu_{ri} + \nu_{si} + \nu_{crs} + \nu_{cri} + \nu_{rsi} + \nu_{csi} + \nu_{crsi,e}$$

where, following the terminology used in Brennan (2001a), μ is the grand mean for the sampled universe and each υ designates an uncorrelated effect, representing a deviation score for its subscripted source of variation (e.g., $\nu_r = \mu_r - \mu$). Such a G study design is represented as $c \times r \times s \times i$. Readers can refer to Figure A-1 in Appendix 9 for the G study design represented by Venn diagram. Here I assume that all the facets in the model are random.

The total observed variance of the scores given by equation (1) over the universe of admissible concept evaluations can be decomposed into fifteen independent variance components shown in equation (2):

(2)

$$\sigma^2(X_{crsi}) = \sigma_c^2 + \sigma_r^2 + \sigma_s^2 + \sigma_i^2 + \sigma_{cr}^2 + \sigma_{cs}^2 + \sigma_{ci}^2 + \sigma_{rs}^2 + \sigma_{ri}^2 + \sigma_{si}^2 + \sigma_{crs}^2 + \sigma_{cri}^2 + \sigma_{csi}^2 + \sigma_{rsi}^2 + \sigma_{crsi,e}^2$$

The variance component for concepts (σ_c^2) quantifies how much evaluations of concepts differ, after averaging over respondents, situations and items. Similarly, the variance component for respondents (σ_r^2) quantifies the extent to which scores for respondents differ, after averaging over concepts, situations and items. The variance component for situations (σ_s^2) quantifies the extent to which scores for situations differ, after averaging over concepts, respondents and items. The variance component for items (σ_i^2) quantifies how much the scores for items differ, after averaging over concepts, situations and respondents. The variance component for a two-way interaction of concepts by respondents (σ_{cr}^2) quantifies the extent to which the relative evaluation of concepts changes from one respondent to another, averaging over situations and items. The variance component for the two-way interaction of concepts by situations (σ_{cs}^2) quantifies the extent to which the relative evaluation of concepts differs from one situation to another, averaging over respondents and items. The variance component for the two-way interaction of concepts by items (σ_{ci}^2) quantifies how much the concepts are ordered differently on different items, averaging over respondents and situations. The variance component for the two-way interaction of respondents by situations (σ_{rs}^2) quantifies how much the evaluation standard of the respondents changes from one situation to another, averaging over concepts and items. The variance component for the three-way interaction of concepts by respondents by situations (σ_{crs}^2) reflects the variability in concepts by respondents by situations interaction, averaging over items. Similarly one can interpret the variance components for other two-way and three-way interactions. Finally, the residual variance component ($\sigma_{crsi,e}^2$) reflects the variability arising from the four-way interaction between concepts, respondents, situations, and items confounded with unmeasured sources of variation.

The purpose of a G study is to obtain estimates of variance components associated with a universe of admissible observations. These estimates can be used to design efficient measurement procedures for operational use and to provide information for making substantive decisions about objects of measurement (usually concepts) in various D studies. The most important D study consideration is the specification of a universe of generalization, which is the universe to which a decision maker wants to generalize based on the results of a particular measurement procedure. Assuming the primary managerial purpose of such a concept test is to assess the relative attractiveness of some concepts, concepts become the object of measurement and the universe of generalization contains all the respondents, items and situations in the universe of admissible observations. In such a decision study, interest focuses on ***mean scores for concepts***, rather than single concept-respondent-item-situation observations that are the focus of G study estimated variance components. This emphasis on mean scores is highlighted by the use of uppercase letters for the facets in the D study $c \times R \times I \times S$ design. The linear model for an observable mean score over n'_r respondents, n'_i items and n'_s situations can be represented as:

(3)

$$X_{cRIS} = \mu + \nu_c + \nu_R + \nu_S + \nu_I + \nu_{cR} + \nu_{cS} + \nu_{cI} + \nu_{RS} + \nu_{RI} + \nu_{SI} + \nu_{cRS} + \nu_{cRI} + \nu_{RSI} + \nu_{cSI} + \nu_{cRSI,e}$$

The variances of the score effects in Equation (3) are called D study variance components. When it is assumed that the population and all facets in the universe of generalization are infinite, these variance components are random effects variance components. They can be estimated using the G study estimated variance components in Equation (2). For example,

$$\sigma^2(R) = \sigma^2(r) / n'_r;$$

$$\sigma^2(S) = \sigma^2(s) / n'_s;$$

$$\sigma^2(I) = \sigma^2(i) / n'_i;$$

$$\sigma^2(cR) = \sigma^2(cr) / n'_r;$$

$$\sigma^2(RS) = \sigma^2(rs) / n'_r n'_s;$$

$$\sigma^2(cRS) = \sigma^2(crs) / n'_r n'_s;$$

$$\sigma^2(IRS) = \sigma^2(irs) / n'_i n'_r n'_s;$$

$$\sigma^2(cIRS) = \sigma^2(cirs) / n'_i n'_r n'_s.$$

Relative error is shown by:

(4) $\delta_c = \nu_{cI} + \nu_{cS} + \nu_{cR} + \nu_{cSI} + \nu_{cRI} + \nu_{cRS} + \nu_{cRSI,e}$

Note that the relative error variance is the sum of the variance components for the seven effects in Equation (4). Then the G coefficient for concepts is shown by $E\rho^2 = \dfrac{\sigma^2(c)}{\sigma^2(c) + \sigma^2(\delta_c)}$

If the concept facet is the object of measurement, maximizing the discrimination between concepts can be achieved by selecting a number of conditions for each measurement facet (respondents, items, situations) to be included in the D study on the facet of generalization so as to achieve the targeted level of error of measurement. For example, any D study design, such as a single item, a single test situation, and fifty respondents, can be evaluated to determine whether it is expected to be sufficiently reliable to make managerial decisions, or whether requiring more respondents to attend the test, or using multiple items to evaluate the concepts, or testing in more than one situation is necessary. This approach can also identify if particular items or types of respondents perform better in discriminating between concepts.

If the managerial purpose is to identify the types of respondents who are most interested in particular concepts, the interaction between concepts and respondents is the object of measurement. The G coefficient for concepts by respondents is shown by:

$$E\rho^2 = \frac{\sigma^2(cr)}{\sigma^2(cr) + \sigma^2(\delta_{cr})}$$

where $\delta_{cr} = \nu_{crI} + \nu_{crS} + \nu_{crSI,e}$.

4.2 Conceptual Framework and Implicit Assumptions

Below I briefly discuss the four types of factors that contribute to observed variation of concept scores and identify the assumptions that currently seem to be made about them by users of concept testing. Figure 4-1 provides a schematic overview of the framework.

(Insert Figure 4-1 about here)

Concept related factors

The idea of the concept itself is usually thought to determine the response in concept testing. In the conceptual framework presented here, concept related factors define what is to be assessed, including the product category, the type of innovation, and stated product design. Thus they include any attribute levels, features, prices, brand names, positioning and promotions tested in a concept test. By testing multiple concepts, users of concept testing, including conjoint analysis, implicitly assume concept related factors account for a large proportion of variance.

Assumption 1: *Concept related factors are a major contributor to response variation in concept testing.*

But, as in any measurement task, it would be surprising if concept test results did not also reflect the many other non-concept related aspects of the testing procedure.

Response task factors

To execute a concept test, a number of response task factors need to be specified. Tests can vary the number of concepts being evaluated, their order of presentation, their form of representation, the response scale items used, and the response format selected for the items. The survey of new product managers found considerable variation in the design of response tasks (Please refer to Chapter 3 for details). For example, concepts are presented in many forms, most often as stripped descriptions (used in 38% of the tests) and stripped descriptions with visual representation (31%). Numerous structured response items are in use, not just the best-known purchase intent measure. Traditionally, researchers try to control for variation in response task factors by choosing a fixed level of each factor for all the concepts being tested (e.g., same form of representation, same item response format). This control approach only makes sense if the response task factors have little or no effect on the reported response. If a response task factor contributes substantially to response variation, the alternative of sampling over it would be more appropriate.

Assumption 2: *Response task factors make trivial contributions to response variation in concept testing.*

Situational factors

Situational factors include testing occasions, competitive products, and market environments. For example, the same concepts are sometimes tested in multiple locations to generalize the results to the whole market. A number of researchers have found situational variables moderate product evaluations, choice patterns and purchase behavior (e.g., Kakkar and Lutz 1975; Dickinson and Wilby 1997; and Miller, Bruvold and Kernan 1987). However, practitioners rarely test the same concepts on multiple occasions or market environments, even though they wish to generalize their test results from the time of testing to the time and market of potential introduction.

Assumption 3: Situational factors are a minor contributor to response variation in concept testing.

Respondent factors

Concept testing employs a variety of screening criteria to select respondents, with the most commonly used being product class and specific product usage, followed by lead user criteria, influential/market maven, and demographics (Please refer to Chapter 3 for details). Respondent selection is recognized to be an important issue in concept testing, although it may be more straightforward when testing minor innovations, which are similar to existing products or services, than for major innovations. Moreover, the consumer behavior literature has identified a number of characteristics on which individuals differ (e.g., innovativeness, consumer expertise, knowledge level) that could influence how respondents react to new products in concept tests. Although few respondent characteristics have been investigated empirically, it is clear concept testing samples respondents much more than any other factor.

Assumption 4: Respondents are a major contributor to response variation in concept testing.

Cross factor interactions

Consumer researchers recognize that significant product-situation interactions indicate that consumers have stronger or weaker preferences for a given product in different situations (Srivastava 1980; Warshaw 1980). Similarly, segment-product and segment-situation interactions can be significant predictors in the case of frequently purchased consumer goods, consumer durables, industrial goods or services (Green and DeSarbo 1979; Leigh and Martin 1981). Thus, concept by situation, concept by respondent, and respondent by situation interactions are of potential concern in concept testing.

When conducting concept tests, users would like all cross factor interaction effects to be negligible. One exception is the variance component due to concepts by respondents, which is of managerial interest, as a large value indicates that the appeal of a concept varies substantially by segment of respondents. Concepts by respondents interaction can even be the object of measurement in concept testing when marketers need to identify the concepts and the consumers, which have the best "fit". Sometimes identifying the fit is more important than identifying which concepts are more attractive or which respondents are more interested in all the concepts because every consumer has individual needs, preferences, resources and behaviors. Since it is virtually impossible to cater for every customer's individual characteristics, marketers group customers into market segments and develop an appropriate marketing mix for all the consumers in the target segment.

Assumption 5: Concepts by respondents interaction, which indicates segment effects, is a major contributor to response variation in concept testing. Other interaction effects are trivial contributors to response variation in concept testing.

Commercial concept tests are generally conducted using a fixed control condition for some response task and situational factors due to the time and cost constraints. However, as in any measurement procedure, there are always other uncontrolled factors (e.g., time of day of data collection, respondent interviewer) whose effects have been uninvestigated and so are left in the residual error. Then one can assume that:

Assumption 6: Residual error is a major contributor to response variation in concept testing.

Chapter 5 Secondary Data Studies

This chapter reconsiders the design of concept testing from a G theory perspective. From this new perspective, concept testing is a measurement and decision-making process rather than a market research technique. The purpose of concept testing is generalizing, with a known error, from a planned set of customer test responses to a defined universe of generalization consisting of the conditions under which the product could be marketed once developed.

This chapter addresses the design questions:

(1) How many respondents are needed to reliably scale concepts?

(2) Is it worth collecting multiple responses rather than relying on a single response item?

(3) Is purchase intent consistently the best response measure to use in concept testing?

(4) What is gained by sampling respondents from multiple locations?

To answer these questions, G theory is applied to four sets of existing concept testing data to see what sources of variance have been sampled and to obtain estimates of their importance in some common testing scenarios. Table 5-1 summarizes the details of the four secondary datasets. The first dataset, from an academic study of heterogeneous new concepts, was obtained from Lees and Wright (2004). The second dataset was collected commercially for a FMCG (fast-moving consumer good) company in China. The third and fourth datasets, provided by a US marketing research firm, were from concept tests of innovative non-consumer products. These secondary datasets are used to determine whether the implicit assumptions presented in Chapter 4 are correct, and to provide a better idea of the psychometric quality of data collected in concept testing studies. The observed variance components should also enable some conclusions to be drawn about design issues in concept testing.

Table 5-1 Details of Four Secondary Datasets

Secondary Datasets	Academic data	Industry data I	Industry data II	Industry data III
Data Source	Lees and Wright (2004)	Commercial project from a company in China	Commercial project from a US marketing research firm	Commercial project from a US marketing research firm
Types of concepts	Consumer products	Fast-moving consumer goods	Radically new non-consumer products	Radically new non-consumer products
Context	Heterogeneous concepts tested in general population	Similar concepts with small changes tested in product class users	Concepts in the same product category tested in product class users	Concepts in the same product category in product class users
Factors sampled and sample size	Concepts (5), respondents (300), formulations (3) and items (7)	Concepts (34), respondents (90), cities (2), items (5)	Concepts (3), respondents (155), segments (8), items (1)	Concepts (3), respondents (151), item (1)

5.1 Academic Data

I first reexamine data for five heterogeneous new concepts made available by Lees and Wright (2004). The four facets of variation they sampled are ***concepts, respondents***, and two task factors, ***formulations*** and ***evaluation items***. The concepts evaluated were a spray-on hand cleanser, a mint-flavored baking soda toothpaste, a spin fryer, a disposable cell phone and a DVD recorder, selected to include durable and consumable products, high and low priced products, and highly innovative products and line extensions. The respondents were registered voters in New Zealand. The three concept formulations investigated were a stripped description, an embellished description, and a visual representation. The evaluation items were five seven-point rating scales asking about problem-solving ability, believability, uniqueness, likelihood to tell, and likelihood to recommend, plus the popular five-point purchase intent scale and an eleven-point purchase plan in the next twelve months scale. The concept testing data were collected using a split-sample mail survey with the three concept formulations as the treatments. The five concepts were tested using each formulation, with voters providing their responses to all five concepts. From a G theory perspective, respondents are nested within concept formulations and crossed with concepts and with items (Refer to Appendix 9 Figure A-2 for the Venn diagram).

Here I assume the study randomly sampled from a large population of conditions for each of the four facets and use GENOVA (Crick and Brennan 1983) to estimate the variance components for the eleven estimable main and interaction effects shown in Table 5-2. To aid in interpretation the table also shows the upper and lower bounds for the 95% confidence interval for each estimate (following Burdick and Graybill 1992) and the percent of variance due to each source. The percent of variance due to concepts is not large, at a modest 5.3%, indicating that other effects contribute most of the variance. The variance due to formulations and concepts by formulations are both estimated to be negative, and so are treated as zero[2]. The former indicates that the particular formulation

---

[2] By definition, variance components are nonnegative. But sometimes sampling variability may cause negative estimates. Using EMS procedure (Cronbach's strategy), I simply set all negative estimates to zero. Note that other estimated variance components may be biased. However, the EMS procedure might be preferable if there were a substantive, theory-based reason to believe that the variance component associated with the negative estimate is indeed zero. An alternative procedure is to use Brennan's algorithm, which keeps the actual values of the negative estimates for verification purposes, then other estimated variance components are not biased (Brennan 2001a, p.84-85).

used does not contribute to the apparent level of interest in the concepts. The latter suggests that the relative attractiveness of concepts is not affected by the formulation used. The variance due to items (8.9%) indicates that different items generate different average scores. These conclusions are consistent with those of Lees and Wright (2004). However, the interaction effect between concepts and items (5.0%) suggests that the concepts are ordered somewhat differently on different items. The variance due to respondents cannot be estimated separately because of the nesting, but the confounded effect is consistent with it being large. The large interaction effect between concepts and respondents (20.2%) indicates that respondents differ substantially in their interest in the five concepts. The response to a particular concept varies a lot depending on the chosen segment of respondents.

Table 5-2

Variance Estimates and Percent of Variance for the Academic Data

Sources of variance	Variance component	Estimate	Standard error	Lower bound[b]	Upper bound[b]	%
1. Concepts (C)	$\sigma^2(C)$	.134	.093	.039	.786	5.3
2. Formulations (F)	$\sigma^2(F)$	-.003	.001	-.079	.287	0.0
3. Respondents within F (R:F)	$\sigma^2(R:F)$	.296	.041	.235	.370	11.7
4. Items (I)[a]	$\sigma^2(I)$	.225	.146	.079	1.247	8.9
5. CF	$\sigma^2(CF)$	-.004	.002	-.007	.005	0.0
6. CR:F	$\sigma^2(CR:F)$	.512	.029	.466	.563	20.2
7. CI	$\sigma^2(CI)$	.128	.044	.076	.258	5.0
8. FI	$\sigma^2(FI)$	.000	.003	-.003	.010	0.0
9. IR:F	$\sigma^2(IR:F)$	.253	.019	.223	.286	10.0
10. CFI	$\sigma^2(CFI)$	.005	.004	.001	.013	0.2
11. Residuals	$\sigma^2(CIR:F,e)$	.986	.020	.954	1.020	38.9

[a] Only five 7-point items are included.

[b] Upper and lower bounds for 95% confidence interval for the estimate.

Table 5-3 lists the implicit assumptions made about the contribution design factors make to concept testing variance. The third column in the Table 5-2 summarizes whether the observed variance components for the academic data reported in Table 5-1 are consistent with those assumptions. Supported are the assumptions that formulations, a task factor, make a trivial contribution and that respondents, concepts by respondents, and residual error make major contributions. Not supported are the assumptions that concepts make a major contribution and that items, the other task factor, and other interactions make trivial contributions. The situational assumption is not tested.

Table 5-3
Assumed versus Actual Contribution to
Variance Made by Sources of Variance in Concept Testing

Design factors that are a source of variance	Assumed contribution	Actual contribution made for a source of data			
		Academic	FMCG	Innovative 1	Innovative 2
1. Concepts	Major	No, modest	No, zero	No, modest	No, zero
2. Response tasks					
- Items	Trivial	No, modest	Yes	-	-
- Formulations	Trivial	Yes	-	-	-
3. Situational					
- Cities	Minor	-	Yes	-	-
- Hospital segments	Minor	-	-	Yes	-
4. Respondents	Major	Yes	Yes	Yes	Yes
5. Interactions					
- Concepts by respondents	Major	Yes	No, modest	*	*
- Other interactions	Trivial	No	No	Yes	-
6. Residual error	Major	Yes	Yes	Yes	Yes

Note - Not sampled from in this design.
* Not separately estimable, as included in residual.

If the primary managerial purpose of a concept test is to assess the relative attractiveness of some concepts, the variance component for the main effect of concepts needs to be large relative to those for all the interactions involving concepts. Here only 5.3% of the variance is due to concepts. However, discrimination between concepts can be increased by sampling more conditions of the facets of generalization (i.e., respondents, items, concept formulations). The usual practice of employing only one concept formulation in a study is supported by the trivial size of all formulation interaction effects that do not involve respondents. Table 5-4 illustrates the effect of increasing the number of randomly chosen respondents and items when comparing concepts in decision studies. If 10 respondents evaluate the concepts on a single item, the expected G-coefficient for concepts is only 0.32. Additional respondents and items are necessary to achieve satisfactory G-coefficient levels for decision-making. For example, 30 respondents evaluating concepts on 10 items or 100 respondents evaluating concepts on 5 items are needed to reach an expected G-coefficient level of 0.80.

In reality, not all facets involved in decision studies are random. The concepts can clearly be considered as randomly sampled from an infinite population. Respondents that are sampled using a local voting register as a sampling frame can also be considered as randomly sampled from a large population. However, most organizations use the same items in all of their concept tests, making items a fixed rather than a random facet. The expected G-coefficient when averaging over five fixed items will be greater than shown for the five random items in Table 5-4. The G-coefficients for the restricted universe of generalization should be larger because the variance due to items by concepts is moved from the error term to become part of universe score variance (Brennan 2001a, p. 99 and p. 122). Table 5-5 compares the G-coefficients when the five items are fixed versus random. For example, the expected G coefficient for 50 respondents is 0.77 if five items are chosen at random but is 0.91 if five items are used as a fixed scale.

To determine which particular item performs the best, I also estimate the variance components separately for each evaluation item. As shown by the percent of variance results in Table 5-6, the problem-solving item discriminates best for scaling concepts, as concepts account for 20.1% of its variance, and it provides the best G coefficients for scaling concepts. Other relatively effective items are purchase intent (15.6%), uniqueness (13.0%) and purchase plan in the next twelve months (11.9%). Believability is the least effective item for scaling concepts. Note that the findings are quite different if the objective is to scale respondents, with likelihood to recommend (33.7%) most effective and problem-solving (14.6%) the least effective item.

To determine whether the results from these academic data are representative, I sought out data from more typical commercial concept testing projects.

Table 5-4

Designed Reduction of Error Variance When Scaling Concepts

G study	Alternative Decision Study Designs									
Number of concepts	5	5	5	5	5	5	5	5	5	5
Number of respondents	10	10	10	30	30	50	50	100	100	150
Number of items	1	5	10	5	10	5	10	5	10	10
Number of formulations	1	1	1	1	1	1	1	1	1	1
Relative error variance	.28	.10	.07	.05	.03	.04	.03	.03	.02	.02
G-coefficient for concepts	.32	.58	.64	.73	.80	.77	.84	.80	.87	.88

Table 5-5

Comparison of G-coefficients When Scaling Concepts

G study	Alternative Decision Study Designs							
	Item Random	Item Fixed	Item Random	Item Fixed	Item Random	Item Fixed	Item Random	Item Fixed
Number of concepts	5	5	5	5	5	5	5	5
Number of respondents	10	10	30	30	50	50	100	100
Number of items	5	5	5	5	5	5	5	5
Number of formulations	1	1	1	1	1	1	1	1
Relative error variance	.10	.07	.05	.02	.04	.02	.03	.01
G-coefficient for concepts	.58	.69	.73	.87	.77	.91	.80	.95

Table 5-6

Comparisons of Percent of Variance for Different Evaluation Items

Sources of variance	Solve	Believe	Unique	Tell	Recommend	Purchase intent	Plan
1. Concepts	20.1	2.6	13.0	8.3	7.9	15.6	11.9
2. Formulations	0.0	0.0	0.0	0.6	0.0	0.0	0.0
3. Respondents (Formulations)	14.6	23.0	24.0	26.9	33.7	18.9	25.5
4. Concepts by Formulations	0.0	0.0	0.5	0.0	0.2	0.0	0.2
5. Residuals	65.2	74.4	62.5	64.2	58.2	65.4	62.4
G-coefficient for concepts[a]	.97	.77	.92	.93	.91	.96	.94

For the design of 100 respondents evaluating each concept using one formulation

5.2 Industry Data for a FMCG

A successful FMCG company with a strong market share and a history of concept testing in China made available a dataset from a recent project with the proviso that no proprietary details would be published. It provided data for 34 concepts evaluated by 90 respondents sampled from two cities (a situational factor) on five key measures, namely purchase intention, perception of newness and difference, perception of price, believability and functions/features importance. All of the respondents assessed all of the concepts on all five measures using zero to ten point scales with labeled end-points.

From a G theory perspective the four-facet data consist of concepts with 34 levels, crossed with respondents with 45 levels, nested within cities with 2 levels, and crossed with items with 5 levels (Refer to Appendix 9 Figure A-3 for the Venn diagram). Table 5-7 reports the estimated variance components, the upper and lower bounds for the 95% confidence interval for each estimate and the percent of variance attributable to each estimable source when treating the four facets as random.

Table 5-7
Variance Estimates and Percent of Variance for FMCG Data

Sources of variance	Variance component	Estimate	Standard error	Lower bound[a]	Upper bound[a]	%
1. Concepts (C)	$\sigma^2(C)$	.000	.009	-.018	.015	0.0
2. Cities (P)	$\sigma^2(P)$	.000	.036	-.280	1.829	0.0
3. Respondents within Cities (R:P)	$\sigma^2(R{:}P)$	.673	.153	.459	.983	9.1
4. Items (I)	$\sigma^2(I)$	.090	.156	-.462	1.044	1.2
5. CP	$\sigma^2(CP)$	.006	.010	-.006	.029	0.1
6. CR:P	$\sigma^2(CR{:}P)$	.467	.037	.408	.529	6.3
7. CI	$\sigma^2(CI)$	.080	.018	.054	.113	1.1
8. PI	$\sigma^2(PI)$	.241	.160	.081	1.352	3.3
9. IR:P	$\sigma^2(IR{:}P)$	1.498	.122	1.315	1.721	20.3
10. CPI	$\sigma^2(CPI)$	.011	.013	-.007	.037	0.2
11. Residuals	$\sigma^2(CIR{:}P,e)$	4.305	.056	4.214	4.400	58.4

[a] Upper and lower bounds for 95% confidence interval for the estimate.

For these data there is no variance at all due to concepts. This indicates that there are no reliable differences among the concepts, so no design can reliably distinguish between levels of evaluation for the population of concepts from which the 34 concepts were drawn. The residual error is very high, accounting for 58.4% of variance. The largest substantive sources of variance are respondents by items, at 20.3%, followed by respondents at 9.1%, and concepts by respondents at 6.3%. Because of the nesting, I cannot separately estimate the effects due to respondents and the interaction between respondents and cities. The variance for the situational factor of cities is zero, suggesting that consumers from different cities had no detectably different average interest in the concepts. These results are summarized in the fourth column of Table 5-3. Supported are the assumptions that items (task factor) and cities (situational factor) make trivial contributions and that respondents and residual error make major contributions. Not supported are the assumptions that concepts and concepts by respondents make major contributions and other interactions make trivial contributions.

As little is known about the concepts, respondents and cities, I can only assume they are random. However, the items could again be fixed rather than random. To again determine which fixed item performs more adequately, I again estimated their variance components separately. As shown by the percent of variance results in Table 5-8, the newness item is best for scaling concepts, accounting for 3.5% of variance, however, this is very low. The other somewhat effective item is believability (2.6%). Price/value and function/feature importance are the least effective items for scaling concepts. In contrast, all items are effective in identifying relative differences between the respondents, especially the price item (46.1%).

Table 5-8

Comparisons of Percent of Variance for Different Evaluation Items

Sources of variance	Purchase	Newness	Price	Believe	Importance
1. Concepts	1.5	3.5	0.9	2.6	0.8
2. Cities	9.1	11.2	14.4	12.3	37
3. Respondents (Cities)	34.3	32.3	46.1	38.0	31.5
4. Concepts by Cities	0.1	0.7	0.5	0.5	1.2
5. Residuals	55.0	52.3	38.1	46.7	62.8

5.3 Data for Related Innovative Products

A US marketing research firm provided access to the third and fourth concept test datasets. The third dataset provided complete concept evaluation data for three radically new products that could be used in hospitals. Physician respondents, who assessed all three concepts using a single five-point purchase intention item, were grouped into eight situational segments based on types of physicians and hospitals. The three facets sampled in the study were concepts with 3 levels, respondents with 155 levels and segments with eight levels. From a G theory perspective, respondents were nested within segments and crossed with concepts (Refer to Appendix 9 Figure A-4 for the Venn diagram).

Variance components are estimated with urGENOVA[3] (Brennan 2001b) since the number of respondents varies by segment, making this an unbalanced design. Table 5-9 reports the five estimable main, the upper and lower bounds for the 95% confidence interval for each estimate and interaction variance components and the percent of variance attributable to each source when treating the three facets as random. For these data there is no variance due to segments, indicating that physicians from different situational segments had no detectably different interest in the concepts. The residual error is very large, accounting for 59.3% of variance. The largest substantive source of variance is respondents, at 34.8%. I cannot separately estimate the effects due to respondents and the interaction between respondents and segments because of the nesting.

Table 5-9
Variance Estimates and Percent of Variance for Innovative Test Data I

Sources of variance	Variance component	Estimate	Lower bound[a]	Upper bound[a]	%
1. Segments (S)	$\sigma^2(S)$	.001	-.013	.033	0.0
2. Respondents within Segments (R:S)	$\sigma^2(R{:}S)$	.170	.123	.231	34.8
3. Concepts (C)	$\sigma^2(C)$	.022	.008	.501	4.6
4. SC	$\sigma^2(SC)$	.006	-.003	.031	1.3
5. Residuals	$\sigma^2(CR{:}S,e)$	.289	.254	.333	59.3

[a] Upper and lower bounds for 95% confidence interval for the estimate.

Presumably the main purpose of the test is to differentiate among the concepts. This means it is desirable for the effect due to the concepts to be large and significant. However, only 4.6% of the variance is due to concepts. To increase discrimination between the concepts in future decision studies requires a large sample size. As shown in Table 5-10, if 100 physicians evaluate the concepts, the expected G coefficient for concepts is only 0.71. Not much is gained by sampling even more respondents, as even 1000 physicians will not reach an expected G-coefficient level of 0.80.

The fourth dataset is similar to the third, but without the respondent segments (Refer to Appendix 9 Figure A-5 for the design represented by the Venn diagram). The data consisted of 151 respondents evaluating three innovative concepts on a single purchase intent item. The variance components for the three estimable sources, namely respondents, concepts and the respondents-by-concepts interaction, are estimated. Table 5-11 reports the three estimable main and interaction variance components, the upper and lower bounds for the 95% confidence interval for each estimate and the percent of variance attributable to each source when treating the two facets as random. The variance component for concepts is zero, again suggesting that no design can reliably distinguish between levels of evaluation for the population of concepts from which the three concepts were drawn. The respondents contribute most of the variance, at 63.8%. The interaction confounded with residual error is also high, accounting for 36.2% of variance.

---

[3] UrGENOVA is used because it can handle unbalanced designs.

Table 5-10
Designed Reduction of Error Variance When Scaling Concepts

G study	Alternative Decision Study Designs							
Number of concepts	3	3	3	3	3	3	3	3
Number of respondents	1	50	100	155	200	250	300	1000
Number of segments	1	1	1	1	1	1	1	1
Relative error variance	.30	.01	.01	.01	.01	.01	.01	.01
G coefficient for concepts	.07	.65	.71	.73	.74	.75	.75	.77

Table 5-11
Variance Estimates and Percent of Variance for Innovative Test Data II

Sources of variance	Variance components	Estimates	Lower bound[a]	Upper bound[a]	%
1. Concepts (C)	$\sigma^2(C)$	-.001	-.001	.004	0.0
2. Respondents (R)	$\sigma^2(R)$	.364	.292	.460	63.8
3. Residuals	$\sigma^2(CR, e)$	.206	.181	.237	36.2

[a] Upper and lower bounds for 95% confidence interval for the estimate.

The results for the innovative concepts are summarized in the fifth and sixth columns of Table 5-3. Supported are the assumptions that situational segments make a trivial contribution and that respondents and residual error make major contributions. Not supported are the assumptions that concepts make a major contribution. The response task assumption is not tested. The concepts-by-respondents interaction is not separately estimable, as it is included in residual error.

5.4 Discussion and Managerial Implications

This chapter treats concept testing as a measurement process used to make managerial decisions. Generalizability analysis of four secondary datasets provides new insights into the implicit assumptions made in the design of concept tests and the psychometric quality of the concept testing data. As summarized in Table 5-3, there is a mixed support for the implicit assumptions made in concept testing. First, the concepts facet is not a major contributor to response variation (i.e., 5.3% in the academic study, 0% in the FMCG data, 4.6% and 0% in the innovative product tests). Second, of the response task factors, concept formulations are a trivial source of variance, but items are not always a trivial source of variance, as shown in the academic data. Third, the situational factors that are investigated are trivial sources of variance (i.e., 0% for cities in the FMCG data, 0.0% for situational segments for the innovative products). Fourth, respondents are always a major contributor to the total variation (i.e., 11.7% in the academic data, 9.1% in the FMCG data, 34.8% and 63.8% for the innovative product tests). Fifth, concepts by respondents are major for the academic data, but only modest for the FMCG data. But other interactions, where estimable, are often not trivial. Finally, residual error is always a major source of variance (i.e., 38.9% in the academic study, 58.4% in the FMCG data, 59.3% and 36.2% for the innovative product tests).

The analyses of the secondary datasets also enable some useful conclusions to be drawn about the four managerial design questions in concept testing. First, because concepts are not a major source of variation in concept testing, relatively large numbers of respondents are required to reliably scale concepts. My survey of practice found the average number of respondents used to evaluate each concept was 92, with smaller numbers for incremental new concepts than radical concepts (Refer to Chapter 3). My G-theory results in this chapter suggest these numbers, which are smaller than those used for other NPD techniques, such as "Voice of the Customer" (Griffin and Hauser 1993), are likely to be too small. In the academic study, the G-coefficient for the design condition when collecting the data using 100 respondents and 5 random items is only 0.80. More than 200 respondents are needed to reach 0.9, the level of generalizability suggested as necessary for managerial decision-making (Finn and Kayande 1997). The conclusion is even less optimistic for two of the industry studies, where the observed variance component for concepts calls into question any ability for the sampled respondents to scale the concepts. In the concept test for innovative products, even 1000 respondents are not sufficient to reach an expected G-coefficient level of 0.80. When the variance component for concepts is zero, as in the FMCG data, no design can reliably scale the concepts and identify which concept is better than another. Concept testing can only be worthwhile if it has other objectives, such as identifying which respondents favor which particular concepts for targeting purposes.

Second, traditional concept testing and conjoint analysis using ratings usually rely on scores on a single item. I found that averaging over items provides considerably more reliable information than relying on a single item. An expected G coefficient when relying on a single item is far lower than when aggregating over 5 items. For example, in the academic study, the G coefficient when collecting the data from 100 respondents and using one random item is 0.47 whereas it increases to 0.80 with 5 random items. If the same 5 items are always used as a fixed scale, the G coefficient is 0.95. In the tests of innovative products, where the evaluation relied on a single purchase intent item, the G coefficients are far from being satisfactory. Thus, it would seem unwise to rely on a single evaluation item when comparing concepts.

Third, if a single item has to be used, which specific item is best is very inconsistent and very context-specific. It depends on the data and research objective. In the academic study, the problem-solving item performs significantly better than the other items when scaling concepts. In the FMCG study, the newness item is the least ineffective of the generally poor items when scaling concepts. The popular purchase intention is never the best single item to use.

Fourth, not much is gained by sampling levels of the response task factor of concept formulations in the academic study, or levels of the situational factors, namely cities and situational segments, investigated in the industry studies.

Finally, concept testing should be designed to meet the needs of specific managerial tasks. In the FMCG study, the variance for concepts is zero, so managers can't expect to reliably identify which concept is better than the rest. However, the variance for respondents is 9.1%, which is big enough to suggest managers could identify these respondents who are more interested in all new concepts. In addition, the variance due to concepts by respondents is 6.3%, suggesting there could be identifiable segments of respondents who are more interested in some of the concepts and managers could segment the respondents according to their level of interests in some of the concepts.

My results were obtained for traditional concept testing. However, I would be surprised if similar results were not obtained for conjoint studies, which can be viewed as testing concepts generated using a factorial design for a set of concept factors, such as attributes and brand names.

5.5 Conclusions on the Design Issues

This chapter demonstrates the value of a new approach to assessing the psychometric quality of concept testing data. The new criteria in the approach are the variance components obtained in a G study for the factors contributed to the response reported in concept testing and the G coefficient for scaling the relevant objects of measurement (usually concepts).

The results of the generalizability analyses of four secondary datasets suggest that the implicit assumptions about sources of variance currently made by the users of concept testing are not all supported. In particular, concepts themselves are not always a major source of variance. In fact, in two out of four sets of data they contribute no variance and in the other two the amount of variance attributable to concepts is small. In addition, response task factors and interactions are not always trivial sources of variance. Items and the two-way interaction between respondents and items make modest contributions in the academic data.

The analyses also illustrate the fact that the appropriate design of concept testing varies with the nature of the managerial task. The academic study is good at scaling concepts, while the industry data could only identify which respondents are more interested in all or particular concepts. The G theory approach helps identify how concept testing can be redesigned to be more effective and more reliable. An important conclusion is that averaging over a set of items provides more generalizable information than relying on a single item. The best single item for discriminating between concepts is dependent on the research situation, but the best item is not the commonly used purchase intent item.

5.6 Limitations of the Secondary Data Studies

A limitation of the secondary data studies is the fact that some sources of variance could not be investigated with any of the secondary data sets. A potentially important source that is not investigated is testing occasions. This is unfortunate, because a time delay between concept testing and potential market introduction is an inherent feature of new product development. Changes in the environment can create unexpected opportunities for or threats to the new product introduction. On the one hand, most unexpected changes make the new product introduction more challenging. For example, RCA's market research in the 1970s assumed that the price of hardware and software for their disc player system would be substantially lower than those for VCRs, and didn't take videocassette rental into account at all (Graham 1986). But, by the time the RCA VideoDisc hit the market in 1981, VCR's were all established, and few consumers wanted a VideoDisc player when for about the same price they could get a VCR that both played and recorded, and they could rent videocassette movies instead of purchasing them (Howe 2004). On the other hand, the changes can create an unexpectedly more favorable opportunity. In 1995 Portugal Telecom far exceeded the market penetration forecasts made for its pioneering MIMO prepaid mobile phone service by enabling cards to be easily recharged at any ATM in Portugal, capitalizing on infrastructure put in place to handle the demands of a prepaid motorway traffic control system (Cavalho 2006).

Sometimes, marketers can modify their marketing mix to negate a negative effect of an environmental surprise. For example, in the aftermath of September 11 2001, GM offered auto purchaser either zero percent financing for up to sixty months or a cash rebate. Due to their popularity, most vehicle manufacturers soon were offering similar incentives on most models, including their newly introduced models (Corrado, Dunn and Otoo 2006). At the same time, while traditional airlines suffered and cut back service, some lost cost carriers saw the aftermath of 9/11 as an opportunity to accelerate their expansion (Jones 2005). In 2002 Britain's easyJet acquired Go Fly, added more services and took advantage of gates vacated by the traditional airlines to introduce new routes, eight from London's Gatwick and four from Paris Orly or Charles de Gaulle.

In practice, product managers have to assume that evaluations generalize from the time (and research environment) of concept testing to the time (and market environment) of market introduction. This clearly requires respondent evaluations to generalize over occasions, but no evidence on the issue was available from the secondary studies, as summarized in Table 5-3. The academic concept testing literature summarized earlier in Table 2-1 provides no help; researchers and practitioners have not reported testing the same concepts on multiple occasions. Therefore, investigating the generalizability of concept testing results over occasions is a priority issue for my primary research.

Another issue not addressable from the secondary data is respondent selection. It is unknown whether some types of respondents provide higher quality evaluation data than others for some types of new concepts. My primary studies will also examine individual difference effects in the generalizability of concept testing.

To investigate these issues, it would be necessary to conduct primary studies to concept test minor and major innovations, over multiple test occasions, using respondents who can be clustered into segments on characteristics assumed to influence their test responses or their predictive capabilities.

Finally, the secondary data studies focus on the internal psychometric analysis of concept testing, while ignoring the question of predictive validity. One could argue that such internal psychometric analyses could lead product managers to take unfortunate actions that improve the internal validity of concept testing at the expense of external validity. For example, there will be more variance due to concepts in concept testing G study when sampling from a population that includes very bad concepts, but their inclusion will be counter-productive in helping identify the best out of a set of good concepts. Another example is that we know that employing an index of five items increases generalizability, but we can't tell whether it performs as well as or worse than a single item, such as the popular purchase intention item, for predictive validity. It would be useful to have data for an external market success criterion in a primary study, and examine whether the one item (or the index of items) that discriminates best also predicts best. Future research should try to obtain data for an external validation criterion, such as how the products actually do in the market place, or even in subsequent product placement tests. I will return to this topic in Chapter 6.

Chapter 6 Primary Studies

To deal with the measurement issues that are not addressable with the secondary data, I conduct a primary study to incorporate concept testing of (1) minor and major innovations, (2) over multiple test occasions, and (3) using respondents who can be clustered into segments on characteristics assumed to influence their test responses or their predictive capabilities. The primary study is important because the main and interaction effects of these three facets, namely concept newness, occasions and respondents, need to be well understood to use concept testing appropriately, but they have seldom been investigated in the concept testing literature (Please refer to Section 6.1 for what research has been done on the relevant topics).

The primary study uses a three-wave web-based experiment to concept test both minor and major innovations over multiple occasions. This chapter first presents the theoretical considerations on the research questions to be addressed in the primary study. Next I present the method used to collect the primary data. Then I report the findings for the temporal stability (or generalizability) of concept testing results over occasions and for individual differences in the generalizability of concept testing. I revisit the design issues about evaluation items using the primary data. At the end of the chapter, I report the results from some validation studies.

6.1 Theoretical Considerations

Temporal Stability or Generalizability of the Results of Concept Testing

A time delay between concept testing and potential market introduction is an inherent feature of new product development. Marketing researchers and practitioners have to assume consumer evaluations of products/concepts generalize over occasions. However, little is known about the temporal stability or generalizability of the results of concept testing over time. Rarely have concept-testing studies incorporated more than one occasion (See Table 2-1). There are at least two reasons for this shortcoming. First, doing so is logistically difficult and costly. Second, in operational settings, collecting data on two occasions is not considered a necessary part of the testing process. Thus the following three research questions are formulated:

1. Will concept-testing results generalize over testing occasions sufficiently well to enable managerial conclusions to be drawn from a single occasion test?
2. Are there differences in generalizability for minor and major innovations?
3. Does the generalizability of concept testing differ by decision-making purposes?

To answer these questions, I investigate the importance of occasion as a source of error variance in estimates of the generalizability of concept test scores for both minor and major innovations.

Consistency of preferences over time

Time inconsistent preferences have been investigated in different areas in the behavioral sciences (e.g., dynamic inconsistency of behavioral decision-making in Thaler 1981; delay of gratification in Mischel, Shoda and Rodriguez 1989; self-control in Rachlin 1995; and temporal construal in Liberman and Trope 1998 and Trope and Liberman 2000, 2003). Although most research explains temporal shifts with affective mechanisms (Ainslie and Haslam 1992; Loewenstein 1996), recent theories have focused more on cognitive processes (Trope and Liberman 2003).

Temporal Construal Theory (Trope and Liberman 2000, 2003) proposes that temporal distance changes people's responses to future events by changing the way people mentally represent those events. People tend to focus on concrete aspects of near-future events and abstract aspects of distant-future events. This shift in consideration has been shown to lead to temporally inconsistent preferences. In a concept-testing context, such preference inconsistency would manifest itself as responses to the new concepts that change as the time to market availability changes.

A number of researchers have examined how temporal changes influence consumers' evaluations of the new product and expectations about the product performance. Drawing on the temporal construal theory, Ziamou and Veryzer (2005) demonstrate that the weight consumers place on the functionality and the interface of a new product is a function of the temporal distance (i.e., time for the purchase or use occasion). Specifically, the functionality of the product is valued more in distant future events while the interface of the product is more important in the near future. Malkoc and Zauberman (2006) examine how temporal framing (deferring versus expediting) of a decision will lead to different degrees of bias towards the present and find a greater decline in consumers' discount rates with time horizon when deferring than when expediting consumptions. Monga and Houston (2006) demonstrate the fading optimism in products and show that confidence about the product performance drops when performance is about to be revealed – soon after choice, or some time later. In particular, they postulate cognitive dissonance and strategic management effects that could change expectations of performance over time. However, these mechanisms seem to be tied to the choice of a product triggering or causing the change. Concept testing does not itself involve a choice process, but this fading effect could be a more general phenomenon that doesn't necessarily require a choice to have been made.

On the basis of Monga and Houston (2006), consumers' evaluations of concepts will drop over occasions because the optimistic expectations about product performance will fade away. This could also be explained by temporal construal theory if subjects focus on more concrete aspects on a second exposure to concepts. On the other hand, the mere exposure effect suggests that simply exposing experimental subjects to a picture or a piece of music briefly led those subjects to later rate it more positively than other, similar stimuli which they had not been shown earlier (Zajonc 1968). On the basis of this effect, consumers' evaluations of concepts will increase over time. These are both predictions for changes in concept test means over time. Hence I have the following two contrasting hypotheses:

Hypothesis 1a: Mean concept scores increase over testing occasions.

Hypothesis 1b: Mean concept scores decrease over testing occasions.

Occasion effects for minor and major innovations

Educational researchers investigating the stability of performance assessments (Ruiz-Primo, Baxter and Shavelson 1993; Shavelson, Baxter and Gao 1993; Webb, Schlackman and Sugrue 2000; Brennan 2000) have found variance attributable to the interaction of person, task, and occasion to be very large. However, they find the variance attributable to occasions and the interaction between persons and occasions to be quite small, which indicate that there is no distinguishable variation in performance levels over time and persons are ranked the same over time when averaging over tasks. However, the large persons by tasks by occasions effect indicates that persons perform tasks differently on different occasions, and this could also be the case for the concept testing of minor and/or major innovations, which could be viewed as different tasks.

There has been a long running debate over the applicability of concept testing techniques to innovative new products (Tauber 1974). It is generally accepted that concept testing predicts trial more accurately for minor innovations than it does for radical innovations (Tauber 1974; Hoeffler 2003; Gourville 2005). Hoeffler (2003) suggests that consumers have greater uncertainty when predicting or estimating the benefits of a really new product compared with those of an incrementally new product. Despite some useful work on information acceleration (Urban, et al. 1997), the keys to successful use of concept testing of radical innovations are not well understood. The concept testing of major innovations is less common (See Chapter 3 for details) because there are fewer of them and there are difficulties encountered in translating traditional techniques of concept testing to these nontraditional settings (Hoeffler 2003).

Minor innovations require little change in usage behavior, so consumers are more likely to focus on concrete aspects of the innovation and more likely to follow through on stated intentions to acquire minor innovations (Alexander, Lynch and Wang 2006). Thus consumers will provide more stable evaluations. For more innovative concepts, consumers are more likely to focus on abstract aspects of the innovation, and the times of testing and market entry are further apart and less likely to be exchangeable. This difference exists likely because more innovative concepts are like distant-future events about which people have greater uncertainty. If preferences must be constructed for abstract aspects rather than retrieved for concrete aspects the consumer knows already, the trade-offs elicited are unstable and easily changed by small changes in the measurement context (Fischoff 1991; Payne, Bettman and Johnson 1992; Payne, Bettman and Schkade 1999; Slovic 1995). Evaluation strategies may change over time when consumers confront major innovations. Thus I predict that:

Hypothesis 2a: *The three-way interaction among subjects, concepts and occasions is a substantial contributor to variation in concept testing of both major and minor innovations.*

Hypothesis 2b: *The three-way interaction among subjects, concepts and occasions is more substantial for major innovations than minor innovations.*

Occasion as a hidden facet of measurement

A facet is hidden when variance components are estimated using a data collection design which does not explicitly sample conditions of the facet, creating interpretational complexities and potential bias in generalizability statistics (see Brennan 2001a, p.149). Occasion remains a largely unexplored potentially hidden facet in marketing applications (Finn 2006). Brennan (2001a) warns that if the interest is really in generalizing over occasions but the data are collected on a single (fixed) occasion, G coefficients (generalizability) will be overestimates because of the confounding of the variance components associated with interactions involving occasions. Therefore I hypothesize that:

Hypothesis 3: Failure to recognizing occasions as an explicit source of variance in the generalizability analyses will lead managers to overestimate the generalizability of their decision studies.

However, the impact of neglecting occasions may vary by purpose of measurement and associated objects of measurement.

Individual Differences in the Generalizability of Concept Testing

The consumer behavior literature has identified several individual characteristics that could influence how subjects respond to new products in concept tests. Few of these characteristics have been thoroughly investigated and little academic research attention has been given to subject selection in concept testing (Klink and Athaide 2006).

Von Hippel (1986) maintains that lead users are a preferred source of input for the development of very new products. Reidenbach and Grimes (1984) show that high knowledge groups provide more accurate evaluations. Duke (1994) agrees that low scores on innovative products might occur if a wide cross section of customers is used in the tests instead of using only innovators to get an early indication of acceptance. Schoormans, Ortt and Bont (1995) suggest that the expertise of consumers would enhance their ability to evaluate both major and minor innovations. Schindler, Holbrook and Greenleaf (1989) suggest that subjects high on innovativeness should be used in concept testing, especially for major innovations. However, Moreau, Lehmann and Markman (2001) conclude that experts are not more prone than novices to adopt discontinuous new products because experts' entrenched knowledge is related to lower comprehension, fewer perceived net benefits, and lower preferences compared with that of novices. More recently, Klink and Athaide (2006) suggest that adoption orientation, as an individual difference variable, should be accounted for in concept testing.

Subject selection issues may be more straightforward when testing minor innovations, which are similar to the existing products or services, rather than major innovations. But major innovations, which confront the consumer with critical tradeoffs or necessary changes in consumption patterns, create a more dissonant decision situation. Tauber (1974) long ago suggested that a major innovation requires screening criteria that reflect its protracted diffusion process.

Practitioners employ a variety of screening criteria to select respondents. The survey of new product managers found the most commonly used screening criteria are behavioral, namely product class usage and specific product usage. Also widely used were lead user (particularly for more radical innovations), influential/market mavens, and demographics. Surprisingly, innovativeness was not a very commonly used criterion. Readers can refer to Chapter 3 for details.

Depending on the nature of the concept being tested, some individuals may be more able to provide reliable responses. Thus three research questions for this portion of the study are:

1. Are the personality traits of the potential customers who are sampled in a concept test an important determinant of concept evaluation results?
2. Do some types of respondents provide substantially higher quality data in concept testing?
3. Do the answers to these two questions change for major versus minor innovations?

To address these questions, I test both minor and major innovation concepts on multiple occasions, using consumers, who can be clustered into segments on consumer characteristics assumed to influence their test responses. Data quality is determined using Generalizability theory, where higher quality means a G coefficient closer to one.

Concept testing assumes that all respondents can understand the concepts and provide an unbiased response to the proposed product (Reidenbach and Grimes 1984). But a number of individual characteristics could influence how accurately subjects can respond to new products. This research investigates personality traits that could influence concept evaluation scores and help identify respondents who provide better quality concept testing data.

Despite a rebirth of enthusiasm for personality trait research in psychology, generated by the five-factor model (McCrae and Costa 1987), consumer personality research has been in the doldrums for decades (Baumgartner 2002). Work employing the hierarchical model of the influence of personality traits on marketing outcomes is just beginning (Brown, et al. 2002). Work on consumer related behavior is also scarce. Therefore, I consider only established consumer research scales, despite some concern about their limitations. I consider scales in the areas of innovativeness, change seeking and cognitive effort for their apparent theoretical relevance and managerial applicability. For example, for a scale to be used to screen respondents, ideally it should be short, easy to administer, and have proven reliability in general population samples.

Innovativeness

Consumer innovativeness reflects the tendency to learn about and adopt innovations (new products) within a specific domain of interest (Goldsmith and Hofacker 1991; Manning, Bearden and Madden 1995). Midgley (1977, Ch. 8) long ago argued that innovators should be used for concept testing, and Zaltman and Wallendorf (1979) presented a model of individual resistance to innovations, which Finn (1985) proposed should be used to choose subjects for concept testing.

Consumers who are high on innovativeness seek out new products and are less resistant to more radical new things. More innovative people are more likely to have product expertise in the specific domain, which allows them to understand product information faster, to fill in missing information, to learn more easily, to discriminate between important and unimportant aspects of a product, and to better infer benefits from a product's physical attributes. Therefore, they are more likely to provide better discrimination between new concepts, especially for major innovations. I hypothesize that:

Hypothesis 4a: *Consumers higher on Innovativeness scales will report a higher mean evaluation of concepts.*

Hypothesis 4b: *Consumers higher on Innovativeness scales will provide higher* *quality data (a higher G coefficient), especially for major innovations.*

In this study, I measured innovativeness using three scales. The first is a 6-item Domain Specific Innovativeness (DSI) scale (Goldsmith and Hofacker 1991). The second is a 2-dimensional consumer innovativeness scale developed by Manning, Bearden and Madden (1995), which measures Consumer Independent Judgment Making (CIJM) and Consumer Novelty Seeking (CNS). CIJM is defined as the degree to which an individual makes innovation decisions independently of others. CNS is defined as the desire to seek out new product information. Another innovativeness related trait is consumers' Desire for Unique Consumer Products (DUCP) (Lynn and Harris 1997). DUCP captures the extent to which consumers hold as a personal goal the acquisition and possession of consumer goods, services, and experiences that few others possess.

Change Seeking

Change seeking is the need for variation in one's stimulus input (Steenkamp and Baumgartner 1995). A substantial body of literature has shown that people with high stimulation needs engage in exploratory behavior to a greater extent than people with lower stimulation levels in order to adjust actual stimulation to their higher optimal levels (Zuckerman 1979). Consumers high on change seeking are more likely to try out new and innovative products, value variety in making product choices, and change their purchase behavior in an effort to attain stimulating consumption experiences (Steenkamp and Baumgartner 1995). Thus I predict that consumers high on Change Seeking scales will give more positive evaluations to the new products, especially the more innovative products, and they are more likely to respond diversely to new concepts because of their stronger desire for exploration. Here is the hypothesis:

Hypothesis 5: *Consumers higher on Change Seeking scales will report a higher mean evaluation of concepts.*

In this study I use three scales to capture change seeking. CSI (Change Seeker Index, Steenkamp and Baumgartner 1995) is a preferred measure of optimum stimulation level. Baumgartner and Steenkamp (1996) developed the Exploratory Buying Behavior Tendency (EBBT) scale to capture people's disposition to engage in two forms of exploratory buying behavior, namely, Exploratory Acquisition of Products (EAP) and Exploratory Information Seeking (EIS). EAP reflects a tendency to seek sensory stimulation through risky and innovative product choices and varied consumption experiences. EIS reflects a tendency to obtain the consumption-relevant information out of curiosity.

In contrast, Preference for Consistency scale (PFC) represents a preference for consistent responding (Cialdini, Trost and Newsom 1995). High PFC respondents are prone to base their responses to incoming stimuli on the implications of existing (prior entry) variables, such as previous expectancies, commitments and choices. Low PFC individuals do not weigh prior entry variables so heavily in their responses; they are open and oriented to the new in ways that are relatively unconstrained by previous standards. However, this should have no effect on the grand mean and G-coefficients. Thus no differential predictions for PFC are made here.

Cognitive Effort

Concept testing assumes respondents will pay attention to detailed information about concepts, engage in intensive information processing, form considered evaluations and report them accurately. In other words, it assumes all respondents will have the same high level of involvement and will be equally motivated to exert the cognitive effort needed at each stage of processing. However, respondents could vary substantially in the cognitive effort they will make to provide their responses. Consumers who are high on cognitive effort are more likely to develop a clear evaluation of a newly encountered product. Thus they should provide higher quality response data than consumers who are lower on cognitive effort, but are no more likely to want the concepts. I hypothesize that:

Hypothesis 6: *Consumers higher on Cognitive Effort scales will provide higher quality data (a higher G coefficient) when scaling concepts, especially for major innovations.*

In this research, the three cognitive effort related scales investigated are involvement, the Need to Evaluate Scale (NES) (Jarvis and Petty 1996) and the Need For Precision (NFP) (Viswanathan 1997). Involvement assesses a subject's interest in and concern about the task he or she is performing (Ozanne, Brucks and Grewal 1992). NES is considered because concept test requires an evaluative response. Individuals differ in the extent to which they chronically engage in evaluative responding. Jarvis and Petty (1996) demonstrated that those who score high on the NES are more likely to form attitudes toward the objects they encounter and may be more likely to engage in evaluative thought about unfamiliar major innovations. NFP captures individuals' differences in their preference for engaging in a relatively fine-grained or precise mode of processing. As suggested by Viswanathan (1997), individuals high on NFP would be more likely to engage in systematic processing by investing cognitive resources in examining large amounts of information, while low NFP people may be more likely to use a subset of available information and simple decision rules.

Social Desirability

Personality research that does not recognize the likelihood of social desirability bias can lead to unwarranted conclusions about consumers' psychological traits (Crowne and Marlow 1960, Paulus 1991). Social desirability is the degree to which people respond in socially acceptable terms in order to gain the approval of others (Richins and Dawson 1992). In concept testing, individuals who are high on social desirability will generate responses that partially reflect beliefs about how others view the stimuli. Thus I would expect their responses to be biased towards the stimuli means. Social desirability should have no effect on the mean evaluations of concepts. But I predict that:

Hypothesis 7: Consumers less susceptible to Social Desirability bias will provide higher quality data (a higher G coefficient) in concept testing.

This research uses the Richins and Dawson (1992) Social Desirability scale.

In summary, the segment effects lead to several hypotheses about individual differences in terms of the mean evaluation of concept stimuli and the generalizability of concept testing. In my primary study, I examine whether a number of personality traits (1) do influence concept evaluation scores and (2) can be used to identify respondents who provide substantially higher quality data in concept testing.

6.2 Method

The study collected concept evaluations of ten consumer appliances from members of an online panel (IOCS, the Institute for Online Consumer Studies) on three occasions, approximately a month apart.

Stimuli preparation

A pool of 20 widely varied new consumer appliances targeted at individuals was sampled from the Internet (i.e., www.appliances.com). I assumed ten of them to be relatively innovative new products and the other ten were less innovative new products (product-line extensions, product improvements and style changes). Concepts, consisting of an image and a paragraph description, were then pretested to confirm which were minor and major innovations, based on perceptions of the potential customers (Blythe 1999). I attempted to keep a consistent concept presentation format, controlling for word quality, word length and graphic style (See the 20 product descriptions in Appendix 2). In addition, all concepts were presented to the respondents without any company or brand identification. Thus, the reactions were to the pure product concept without the influence of the established images or values associated with the company or its brand name.

Pretest

The pretest was conducted via a web-based survey that allowed the concept presentation order to be randomized. A total of 54 participants in the IOCS panel signed up for the pretest. They were asked to assess the newness and clearness of the 20 descriptions of appliances monadically. Please see the Appendix 2 for the newness and clearness measures used in the pretest. My newness scale was adapted from Lee and O'Connor (2003).

The newness scores (the mean of the six newness items) were used to identify 5 appliances that respondents perceived as minor innovations and 5 that respondents perceived as major innovations for the follow-up concept testing generalizability study. Please see Appendix 3 for the detailed newness scores for all 20 concepts. Five appliances with the highest newness scores are selected to be in the group of major innovations while five with the lowest newness scores (except oral care system) are selected to be in the group of minor innovations. Steam iron was selected though its newness score is a bit higher than oral care system, because the latter is quite similar to the dental water jet that ~~has~~ already been selected to the group of minor innovations. I try to include the products with more diversity in each group and avoid the comparisons among the similar products.

The clearness scores for all 20 concepts range from 5.74 to 6.40 out of 7 with an average of 6.06 and a standard deviation of 0.21. It appears that there is no significant difference among the product descriptions in terms of clearness. I consider the clearness scores for all the 20 concepts are acceptable with no remarkable lemons included.

Table 6-1

Newness and Clearness Scores for Minor and Major Innovations

Newness	Concept	Newness Score	Clearness Score
Minor	Ear Thermometer	2.16	6.40
Minor	Stereo Radio	2.59	6.37
Minor	Dental Water Jet	2.70	6.30
Minor	MP3 Player	2.74	6.00
Minor	Steam Iron	2.85	6.28
Minor	*Mean*	*2.61*	*6.27*
Major	Hair Cutting Tool	3.49	6.09
Major	Digital Camera	3.60	5.79
Major	Smartphone 2	3.65	6.02
Major	Portable Media Center	3.87	5.81
Major	Personal Computer	3.93	5.81
Major	*Mean*	*3.71*	*5.90*

Table 6-1 shows the newness scores and clearness scores for the ten selected concepts. The mean newness scores for minor innovations and major innovations are 2.61 versus 3.71, which are significantly different at .001 level (two-tailed paired sample *t*-test was used because the same respondents responded to the two types of products for which the means are being compared). The coefficient alpha for the newness scale is 0.888. A factor analysis of the newness scores suggests a single factor solution (the first eigenvalue is 3.869, remarkably greater than the second eigenvalue of 0.785), accounting for 64.5% of the total variance. These results confirm that the newness scale I used in the pretest is unidimensional.

To confirm the newness difference between minor and major innovations, I use GENOVA (Crick and Brennan 1983) to estimate the variance components for the eleven estimable main and interaction effects for the ten selected concepts. The description of the pretest is not in terms of the traditional experimental design terminology, which applies better to fixed effects models. I describe it from a G theory perspective in terms of the random effects populations, the sampling of conditions (levels), and the crossing or nesting of the factors. The object of measurement is concepts and the three facets of variation I sampled in the pretest are respondents, newness level and items. From a G theory perspective, concepts are nested within newness level and crossed with respondents and newness items (Refer to Appendix 9 Figure A-6 for the design represented by the Venn diagram). Here I assume the pretest randomly sampled from a large populations of conditions for concepts, respondents and newness items. Newness level is considered a fixed factor with two levels, namely major and minor innovations.

Table 6-2 shows the variance estimates and the percent of variance due to each source and the standard error for each variance component estimate. As shown in Table 6-2, the variance (main effect) due to newness is 0.592 that accounts for 22.8% of the total variance, indicating that the two types of concepts differ substantially in their newness.

Table 6-2

Variance Component Estimates for Pretest with 10 Selected Concepts

Modeled Source	Variance Estimate	Standard Error	%
Newness (N)	.592	-	22.8
Concepts within Newness (C:N)	.034	.024	1.3
Respondents (R)	.463	.119	17.8
Items (I)	.109	.062	4.2
NR	.025	.038	1.0
NI	.012	.010	0.5
RC:N	.593	.053	22.8
IC:N	.015	.006	0.6
RI	.168	.022	6.5
NRI	.053	.016	2.0
Residuals	.535	.019	20.6
Total	2.599		100.0

Generalizability study design of the main study

The five variables to be studied in the main study and their associated abbreviations are concept newness to consumers (N), concepts (C), subjects (R), evaluation items (I) and test occasions (O).

1. Concept newness – A fixed facet with two levels, namely major and minor innovations.
2. Concepts – A random facet from which I sample appliances. I have a total of ten appliance concepts with five concepts nested within each level of concept

innovativeness.

3. Subjects – A random facet beginning with 105 different IOCS panelists, sufficient to allow for some attrition over the three occasions.
4. Evaluation items – A random facet consisting of six commonly used concept-testing items expressed as seven-point semantic differential scales. I used categorical rather than continuous scales because the former are most commonly used in the current practice (Refer to Chapter 3). The specific items chosen were (1) Purchase intention, (2) Liking, (3) Importance, (4) Uniqueness, (5) Problem solving, and (6) Believability. The items were coded from 1 to 7 with more positive responses given the higher values. See Appendix 4 for the precise wording of each item. In reality, most organizations use the same items in all of their concept tests, making items a fixed rather than a random facet.
5. Test occasions – A random facet using a test-retest-retest research design with one-month interval between the tests.

Data collection procedure

A three-wave study was conducted between September 30 and December 1, 2005. I recruited subjects from IOCS participant panel via e-mail. In the first wave, 105 respondents evaluated ten appliance descriptions on six concept evaluation items. A sequential monadic design was used, with order of presentation randomized over subjects to minimize order bias. The respondents also completed the Need to Evaluate (Jarvis and Petty 1996), the Need For Precision (Viswanathan 1997) and the Buying Impulsivity (Rook and Fisher 1995) scales. One month and two months later, the same participants were asked to re-evaluate the same ten concepts again, providing data for some respondents on all three occasions. On the second occasion they also completed the Domain-Specific Innovativeness (Goldsmith and Hofacker 1991), the Consumer Independent Judgment Making and Consumer Novelty Seeking (Manning, Bearden and Madden 1995), the Desire for Unique Consumer Products (Lynn and Harris 1997), the Involvement (Ozanne, Brucks and Grewal 1992) and the Social Desirability (Richins and Dawson 1992) scales, and on the third occasion they completed the Exploratory Buying Behavior Tendencies (Baumgartner and Steenkamp 1996), the Change Seeker Index

(Steenkamp and Baumgartner 1994) and the Preference For Consistency (Cialdini, Trost and Newsom 1995) scales. See Appendix 4 for the specific items of the personality trait scales. The respondents were offered $6, $8, and $8 fees for their completed first, second and third wave responses. To encourage retention, a $6 bonus was provided for participants who completed all three waves. This procedure resulted in 78 subjects providing data on all three occasions.

Here let me explain why I sampled three occasions, approximately a month apart. If generalization is intended to a broader set of occasions than the single occasion on which data are collected, then occasions is a random factor. Any single occasion is randomly sampled from a universe of occasions over which a manager would like to generalize the concept test results. When only one occasion is sampled, the facet of occasions is hidden in the sense that variance components associated with occasions are confounded with other variance components and the estimates of the generalizability of concept tests (G-coefficients) will be biased (Brennan 2001). That is why the study sampled "three" occasions. One month was used as an interval because I expected one-month to be long enough for the respondents to forget the details of the product descriptions and the exact responses they had given on a previous occasion.

6.3 The Generalizability of Concept Testing Over Occasions

In this section, I present the data analysis results and conclude with the managerial implications for the generalizability of concept testing over occasions. The results are presented in the following three parts, namely the stability of concept scores, generalizability study results for the multiple occasion data and generalizability of concept scores for different decision-making purposes.

The stability of concept scores

To begin I examined the mean evaluation scores for each of the ten concepts across three occasions. As shown in Table 6-3, one striking observation is that the mean concept scores obtained on the first occasion are significantly higher than those obtained on the second and third occasions. The only exception is for concept 3 on the second occasion.

Concept 1 and concept 10 received remarkably different evaluations across three occasions. Except for concepts 1, 3 and 10, there is no significant difference in mean concept evaluations between the second and the third occasions. This pattern of results is inconsistent with predictions that might be made based on a mere exposure effect (Zajonc 1968). This could be explained if subjects focus on more concrete aspects on a second exposure to concepts. This result provided evidence of instability of the concept scores across occasions, indicating that occasion would be a major source of error variability.

Table 6-3
Mean Evaluation Scores by Concepts and Occasions

Concept		Occasion 1	Occasion 2	Occasion 3
Minor	2	5.41*	5.15	5.12
Innovations	5	5.21*	4.87	4.96
	3	5.10	5.06*	4.85
	4	4.86*	4.65	4.66
	1	4.75*	4.28*	4.42
Major	8	5.50*	5.06	5.15
Innovations	7	5.45*	5.18	5.15
	6	5.33*	5.11	5.16
	9	5.12*	4.75	4.68
	10	4.91*	4.62*	4.46
Mean		5.16*	4.87	4.86

*- $p<0.05$ for occasion 1 compared with occasion 2 or occasion 2 compared with occasion 3, two-tailed paired sample *t*-test is used.

A different way to do the same thing – Repeated Measure Analyses

I have run each of the comparisons (mean concept scores on occasion 1 compared with those on occasion 2, mean concept scores on occasion 2 compared with those on occasion 3) using simple paired sample *t* tests at conventional level 0.05 (as shown in Table 6-3). The data I have can also be considered as a set of repeated measures. I have the same 78 respondents evaluating the same 10 concepts using the same 6 evaluation items on three successive occasions. SPSS GLM-repeated measures procedure can be used to run the comparisons among occasions more easily.

I set up the data for the analysis. Seven variables are identified as subjects with 78 levels, concepts with 10 levels, items with 6 levels, newness with 2 levels, and time 1, time 2 and time 3 score. A within-subject factor "Time" is defined to have three levels (time 1, time 2 and time 3). The purposes of the analysis are (1) to investigate whether the differences among the occasions is significant and (2) to study the differences between major innovations and minor innovations.

The detailed descriptive statistics and the overall analysis of variance are shown in the Appendix 5. Sphericity (all the variances of the differences are equal in the population sampled) is a mathematical assumption in repeated measures ANOVA designs. When this assumption is violated, there will be an increase in Type I errors, because the critical values in the F-table are too small. One could say that the F-test is *positively biased* under these conditions. As suggested by Algina and Keselman (1997), there are two approaches to dealing with violations of sphericity, namely to use a correction to the standard ANOVA tests using the Greenhouse-Geisser correction or Huynh-Feldt correction and to use a different test (i.e., one that doesn't assume sphericity, such as MANOVA). In the following two designs, a significant result of Mauchly's sphericity test indicates that sphericity is violated. Using the Huynh-Feldt correction, the main effect of time was significant in both designs. Here are the highlights of the findings:

1. Mean differences among occasions. There are no between-subjects effects in this design (as shown in the first part of Appendix 5). Using the Huynh-Feldt correction the main effect time was significant. Figure 6-1 suggests that there is a large drop in the mean concept scores from occasion 1 to occasion 2. There doesn't seem to be any decrease in mean scores from occasion 2 to occasion 3.

2. Mean differences between major innovations and minor innovations. Concept newness is a between-subjects factor in this design (as shown in the second part of Appendix 5). The between groups test indicates that there the variable newness is significant, consequently in the graph the lines for the two groups are rather far apart. The within subject test using Huynh-Feldt correction indicates that the main effect of time is significant and both groups get less favorable evaluation from occasion 1 to occasion 2. Also, since the lines are parallel, I am not surprised that there is no interaction between occasion and newness, which means that both groups change the evaluation over time but are changing in the same ways. Figure 6-2 demonstrate the within-subjects time effect and between-subjects newness effect.

Figure 6-1
Profile plot for within-subjects time effect

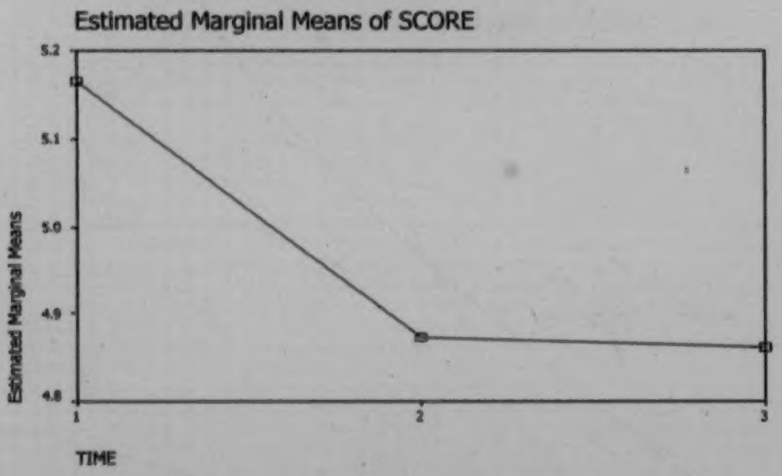
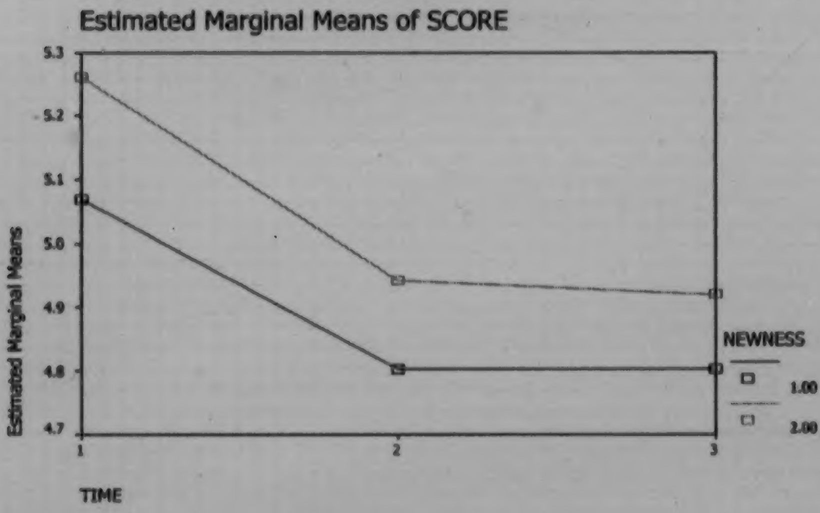


Figure 6-2
Profile plot for within-subjects time effect and newness effect



Generalizability study results for the multiple occasions data

Next I examine the effect of occasions from the G theory variance components perspective. Two contrasting views of the G study design are assumed for analysis. First, occasions is treated as a fixed facet and a concepts nested within concept newness and crossed with respondents and items design is analyzed separately for each occasion. Second, occasions is treated as a random facet and the entire data set is analyzed as a concepts nested within concept newness and crossed with respondents, items and occasions design. The comparison of these two designs makes it possible to assess the importance of accounting for the occasions variation in designing concept tests.

Table 6-4 describes the results of variance component estimates for the first G study design with a hidden occasions facet. To aid in interpretation of the magnitudes, Table 6-4 also shows the percent of variance due to each source. Modeling single occasion data provides estimates for ten main or interaction sources, plus the highest order interaction, which is confounded with residual error. The substantial sources of variance for the Time 1 data are concepts nested within newness by respondents (32.2%), respondents (25.3%) and the highest order interaction confounded with error (19.3%). When modeling the Time 2 data or Time 3 data only, the total variance is bigger over time, but the proportion of variance exhibits a similar pattern.

Table 6-4
Variance Component Estimates for Models with Hidden Occasions

Modeled Source	Time 1		Time 2		Time 3	
	Variance	%	Variance	%	Variance	%
Newness (N)	.000 (-)	0.0	.000 (-)	0.0	.000 (-)	0.0
Concepts within Newness (C:N)	.048 (.029)	1.5	.074 (.041)	2.3	.073 (.041)	2.1
Respondents (R)	.804 (.154)	25.3	.819 (.155)	25.0	.031 (.189)	29.3
Items (I)	.195 (.108)	6.1	.209 (.115)	6.4	.227 (.124)	6.4
NR	.026 (.044)	0.8	.107 (.055)	3.3	.153 (.064)	4.3
NI	.066 (.038)	2.1	.056 (.033)	1.7	.043 (.026)	1.2
RC:N	**1.022** (.064)	32.2	.978 (.062)	29.8	1.025 (.064)	29.2
IC:N	.012 (.004)	0.4	.019 (.006)	0.6	.018 (.006)	0.5
RI	.309 (.027)	9.7	.279 (.025)	8.5	.272 (.024)	7.7
NRI	.083 (.015)	2.6	.089 (.016)	2.7	.098 (.016)	2.8
Residual	.614 (.016)	19.3	.651 (.017)	19.9	.576 (.015)	16.4
Total	3.179		3.281		3.516	

Standard error is given in the parentheses. Number in bold is discussed in the text.

To investigate the hidden occasions facet, estimates for the full 23 sources of variance, including occasions and all its interactions, are shown in Table 6-5. Readers can refer to Appendix 9 Figure A-7 for the full model represented by the Venn diagram. When taking the occasions facet into account, the main effect of occasions and many of its lower order interactions are negligible. However, the variance component for concepts nested within newness by respondents by occasions (0.469, 14%) is quite substantial. Comparing the variance components for the time 1 observations in Table 6-4 with those in Table 6-5 reveals that the hidden occasions facet contributed substantially to overstated variance component values for concepts nested within newness by respondents (i.e., 1.022 vs. 0.539; 32.2% vs. 16.1%).

Table 6-5
Variance Component Estimates Taking Account of Occasions

Modeled Source	Variance estimate	Standard error	%
Newness (N)	.000	-	0.0
Concepts within Newness (C:N)	.066	.035	2.0
Respondents (R)	.711	.021	21.2
Items (I)	.211	.141	6.3
Occasions (O)	.027	.027	0.8
NR	.072	.115	2.1
NI	.056	.001	1.7
NO	.000	.000	0.0
OR	.174	.016	5.2
OI	.000	.000	0.0
RI	.213	.002	6.4
OC:N	.000	.020	0.0
RC:N	**.539**	.039	16.1
IC:N	.016	.032	0.5
NRI	.047	.043	1.4
NOR	.023	.009	0.7
NOI	.000	.022	0.0
ORI	.074	.006	2.2
RIC:N	.173	.004	5.2
ORC:N	**.469**	.001	14.0
OIC:N	.001	.009	0.0
NORI	.043	.007	1.3
Residual	.440	.008	13.1
	3.355		100.0

Numbers in bold are discussed in the text.

As the concepts are grouped into major innovations and minor innovations, I also analyze the concept newness subgroups separately and compare the variance components to identify whether there are differences between minor and major innovations. Table 6-6 compares the variance attributable to the three-way interaction among persons, concepts and occasions for an index of all six items. The proportion of variance due to the three-way interaction of concepts by occasions by respondents (after averaging over six items) for major innovations is more substantial (but not statistically) than that for minor innovations (i.e., 0.487 vs. 0.452; 16% vs. 14%).

To investigate whether a particular item provides more stable evaluation over time, Table 6-6 also reports the same variance components comparison for each (fixed) evaluation item. As shown in Table 6-6, the three-way interaction of concepts by occasions by respondents confounded with error is a substantial contributor to the total variance for each of the six items. It is even more substantial for major innovations than minor innovations except for the case of the uniqueness item. For the importance and purchase intention items, the three-way interaction is significantly larger for major innovations than it is for minor innovations. None of the items provides a substantially more stable evaluation over time than the others.

Table 6-6
Comparison of Variance due to CRO

Response Item	Major Innovations		Minor Innovations	
Purchase intention	1.087*	28%	.906	25%
Liking	1.132	32%	1.053	31%
Uniqueness	.872	37%	1.023*	35%
Problem solving	1.075	28%	.980	26%
Believability	.535	29%	.507	28%
Importance	.945*	39%	.801	31%
Index of six items	.487	16%	.452	14%

* - $p<0.05$; methods discussed in Burdick and Graybill (1992) are used for the significance test. See Appendix 6 for the description and detailed calculation code.

Generalizability of concept scores for different decision-making purposes

As reported in Chapter 3, the survey of new product managers found the most often cited objective of concept testing was to develop further the original idea (cited by 81% of the respondents) where concepts by aspects is the object of measurement. Other widely cited objectives were to estimate the concept's market potential (70%) and eliminate the poor concept (66%), where concepts is the object of measurement; identify the value of concept features (66%), where concepts by aspects is the object of measurement; and help identify the highest potential customer segment (53%), where concepts by segments is the object of measurement. This finding indicates that concepts and concepts by respondents (or segments) are the two most important objects of measurement in concept testing.

To illustrate the potential managerial implications I compare the G coefficients expected in D studies based on the variance components from the full 23 sources model in Table 6-5 with G coefficients based on the hidden occasions variance components from the Time 1 data alone.

First, one primary managerial purpose is to assess the relative attractiveness of the concepts, making concepts the most important object of measurement. Table 6-7 uses the Full Model in Table 6-5 and Time 1 variance components in Table 6-4 to compare the conclusions that would be drawn about the G-coefficients to be expected from the same set of decision study designs when scaling concepts. The results in Table 6-7 show that the expected relative errors are reasonably similar. The G-coefficients for a single random occasion based on the three occasions data were slightly higher than the hidden data. Thus, for this purpose the conclusions drawn about generalizability from data collected on a single occasion would not be problematic. This is just one example. If I use Time 2 or Time 3 data, the results might be different.

(Insert Table 6-7 about here)

A second managerial purpose is to identify the types of respondents who are most interested in particular concepts. Table 6-8 reports results for decision studies for concepts by respondents when practitioners need to identify how respondents respond to particular concepts. Using the data from Time 1 it appears that useful data for segmentation can be obtained by having each concept evaluated by each respondent on one occasion (has a G-coefficient of 0.91), so long as the instrument includes six items. Here the confounding between concepts nested within newness by respondents and three-way interaction of concepts nested within newness by respondents by occasions in the hidden data leads to an overestimated universe-score variance and an underestimated error variance, producing an overestimate of the level of generalizability. Thus the same design in a single random occasion design based on the three occasions data will only provide a G-coefficient of 0.49. To obtain a G-coefficient suitable for applied research (more than 0.9) requires the measures to be collected on at least twenty occasions.

(Insert Table 6-8 about here)

Here I find evidence of bias in variance components estimated when ignoring the existence of a hidden occasions facet. The bias is very substantial for the variance due to concepts by respondents. This component is overestimated because variance due to the three-way interaction of concepts by respondents by occasions is quite substantial, indicating that the relative evaluation of concepts that consumers report varies considerably over occasions. This result violates the assumption of concept testing in which practitioners assume consumer evaluations of concepts to be generalizable over occasions. The effects of the biased estimates are shown to be relatively minor for scaling concepts. The bias effects are substantial for scaling concepts by respondents (i.e., segmenting customers).

Table 6-8
D study Generalizability Coefficients for Concepts by Respondents

	Three Occasions					Time 1		
No. of items	1	6	6	6	6	1	6	12
No. of occasions	1	1	3	5	20	1	1	1
Component								
RC:N	.54	.54	.54	.54	.54	1.02	1.02	1.02
RIC:N	.17	.03	.03	.03	.03	.61	.10	.05
ORC:N	.47	.47	.16	.09	.02	-	-	-
Residual	.44	.07	.00	.01	.00	-	-	-
Relative error	1.08	.57	.19	.14	.06	.61	.10	.05
G for concepts by respondents	.33	.49	.74	.80	.90	.62	.91	.95

Finn (2006) provided an initial demonstration of the consequences of failing to account for the hidden occasions facet of variation in service performance assessment. The present study extends the Finn (2006) research to a new product evaluation research context, demonstrating that the pattern of hidden occasions interaction effects should not be neglected in marketing measurement.

Table 6-7
D study Generalizability Coefficients for Concepts

	Three Occasions						Time 1					
No. of respondents	1	10	10	25	50	100	1	10	10	25	50	100
No. of items	1	1	6	6	6	6	1	1	6	6	6	6
No. of occasions	1	1	1	1	1	1	1	1	1	1	1	1
Components												
C:N	.07	.07	.07	.07	.07	.07	.05	.05	.05	.05	.05	.05
RC:N	.54	.05	.05	.02	.01	.01	1.02	.10	.10	.04	.02	.01
IC:N	.02	.02	.00	.00	.00	.00	.01	.01	.00	.00	.00	.00
RIC:N	.17	.02	.00	.00	.00	.00	.61	.06	.01	.00	.00	.00
OC:N	.00	.00	.00	.00	.00	.00	-	-	-	-	-	-
ORC:N	.47	.05	.05	.02	.01	.00	-	-	-	-	-	-
OIC:N	.00	.00	.00	.00	.00	.00	-	-	-	-	-	-
Residual	.44	.04	.01	.00	.00	.00	-	-	-	-	-	-
Relative error	1.70	.18	.11	.05	.03	.01	1.65	.18	.11	.05	.02	.01
G-coefficient for concepts	.04	.27	.37	.58	.72	.83	.03	.21	.30	.51	.66	.78

Conclusions and Implications

This study set out to examine the generalizability of concept scores across occasions for concept testing of both minor and major innovations. The findings lead to the following conclusions:

Firstly, the results showed a systematic decline in mean concept scores from occasion one to occasions two and three. Mean concept scores are significantly different over occasion 1 versus occasion 2, and occasion 1 versus occasion 3, but not significantly different over occasion 2 versus occasion 3.

Second, the three-way interaction among subjects, concepts and occasions is a substantial contributor to variation in concept testing of both major and minor innovations, and is even more substantial for major innovations than it is for minor innovations. This large interaction effect indicates that consumers evaluate the same concepts differently on different occasions, and this could be a threat to the validity of concept testing.

Third, explicitly recognizing occasions as a facet of error variance influenced the generalizability of the test results. Practitioners of concept testing who allows occasions to remain a hidden facet when designing a concept test may substantially overestimate the generalizability of the data it will collect. However, the extent to which it alters the psychometric quality of the concept test data varies with the nature of the managerial task (e.g., assessing concepts or segmenting customers). The present study provides evidence that it is possible to conceive of a manager implementing the current design of concept testing to scaling concepts; it is hard to conceive of a manager implementing the designs that would be necessary for studies scaling concepts by respondents. To be specific, consumer evaluation of concepts, observed on one occasion in this study is not generalizable, to a whole set of possible occasions (e.g., potential market introduction) if the purpose of the concept testing is to identify the types of respondents who are most interested in particular concepts (segmentation). However, if the purpose of concept testing were to assess the relative attractiveness of the concepts (scaling concepts), the generalizability from data collected on a single occasion would be acceptable.

The results of the hypothesis tests in the primary study are summarized in Table 6-9 at the end of this section. Supported is Hypothesis 2a that the three-way interaction among subjects, concepts and occasions is a substantial contributor to variation for both minor and major innovations. Not supported is Hypothesis 1a that mean concept scores increase over testing occasions. There is mixed support for Hypotheses 1b, 2b and 3. Hypothesis 1b is supported for the data on the first and second occasion except for concept 3, but not supported for the second and the third occasion for the fading effect is very weak from occasion 2 to occasion 3. In particular, mean concept scores obtained on the first occasion are significant higher than those obtained on the second and third occasions at 0.05 level. The only exception is for concept 3 on the second occasion. Concept 1 and concept 10 received statistically different evaluations across three occasions at 0.05 level. Except for concepts 1, 3 and 10, there is no significant difference in mean concept evaluations between the second and the third occasions. Hypothesis 2b that the three-way interaction among subjects, concepts and occasions is more substantial for major innovations than minor innovations is supported for the index of six items and each of the six items except for the uniqueness item. For the importance and purchase intention item, the three-way interaction among subjects, concepts and occasions is statistically larger for major innovations than it is for minor innovations at 0.05 level. Hypothesis 3 that failure to recognizing occasions as an explicit source of variance will lead to the overestimate of the generalizability of concept testing is supported when concept testing is used for the segmentation purpose, but not supported for the purpose of scaling concepts.

Inevitably practitioners will need to generalize consumer evaluation of concepts, observed on one occasion, to a whole set of possible occasions (e.g., potential market introduction). This research provides insight about how well concept testing can generalize over occasions. To increase the dependability of this generalization, more occasions will need to be sampled, even though this will be costly and time consuming. Nevertheless, the importance of concept testing scores in new product introduction decision seems to warrant the expense. As suggested by Finn (2006), managers should conduct at least small scale G studies that sample occasions as a facet of measurement in order to determine the extent to which scores obtained on a single occasion are as generalizable as expected to scores that might be obtained on different, but similar occasions.

Table 6-9
Tests of the Hypotheses Made in the Primary Study

Hypothesis	Decision
H1a - Mean concept scores increase over testing occasions.	Not supported
H1b – Mean concept scores decrease over testing occasions.	Supported for the data on the first versus second occasion, and the data on the first versus third occasion
H2a – The three-way interaction among subjects, concepts and occasions is a substantial contributor to variation in concept testing of both major and minor innovations.	Supported (14% on the data including both types of concepts on three occasions; 16% for major innovations versus 4% for minor innovations)
H2b – The three-way interaction among subjects, concepts and occasions is more substantial for major innovations than minor innovations.	Supported for the index of six items and each of the six items except for the uniqueness item
H3 - Failure to recognizing occasions as an explicit source of variance in the generalizability analyses will lead managers to overestimate the generalizability of their decision studies.	Supported for the segmentation purpose; not supported for scaling concepts
H4a - Consumers higher on Innovativeness scales will report a higher mean evaluation of concepts.	Supported for all Innovativeness scales in testing of major innovations, but not supported for CIJM in testing of minor innovations
H4b - Consumers higher on Innovativeness scales will provide higher quality data (a higher G coefficient), especially for major innovations.	Supported for all Innovativeness scales in all testings, not supported for CIJM in testing of minor innovations
H5 - Consumers higher on Change Seeking scales will report a higher mean evaluation of concepts.	Supported for all Change Seeking scales in testing of major innovations, but not supported for PFC in testing of minor innovations
H6 - Consumers higher on Cognitive Effort scales will provide higher quality data (a higher G coefficient) when scaling concepts.	Supported for NES in all testings and NFP when testing major innovations and minor innovations separately, not supported for NFP when testing all concepts
H7 - Consumers less susceptible to Social Desirability bias will provide higher quality data (a higher G-coefficient) in concept testing.	Supported for SD when testing major innovations, not supported for SD when testing all concepts and minor innovations, not supported for Involvement and EIS in all testings

6.4 Individual Differences in the Generalizability of Concept Testing

In this section, I present the data analysis results and conclude with the managerial implications for the individual differences in the generalizability of concept testing. The results are presented in three parts, namely the dimensionality of the personality trait data, trait segment differences in concept evaluation scores, and finally trait segment differences in the generalizability of concept testing.

Dimensionality of the Personality Trait Data

While it might be informative to begin at the item level and assess to what extent items within particular scales load together, the limited ratio of respondents to total items makes such a strategy impractical (Guadagnoli and Velicer 1988). Therefore the analysis begins with the observed trait data and investigates possible factors underlying the respondent scores. As shown in Table 6-10 the coefficient alpha measures of internal consistency for all the traits except Social Desirability are above the 0.70 standard of acceptability for academic research (Nunnally 1978).

There are various rules to determine the "correct" number of factors in a data set. According to the review from Keeling (2000), eigenvalues of a sample correlation matrix have been used in a variety of ways to make decisions regarding dimensionality. These decisions have commonly been based on the procedures such as Kaiser eigenvalue greater than one rule (Kaiser 1960), Cattell's scree test (Cattell 1966), maximum likelihood test (Jöreskog 1967) and Horn's parallel analysis (Horn 1965). I have looked at the results from different procedures. An example is the scree plot in Figure 6-3. To determine the break point of scree plot, I began at the right of the diagram, and placed the best fitting lines. When I did this, I obtained 3 factors. I also employed principal axis factoring (PAF), which uses squared multiple correlations for the communality estimates, and oblimin transformation to allow for a correlated factor solution. Please refer to Appendix 7 for the detail output. When I examined the factor pattern, 4 of the 13 variables had complexity 2. If I used 0.40 as the cut-off between a salient loading and a non-salient loading, then the number of complex variables is reduced to one (i.e., CNS). The simple structure of the factor loadings in Table 6-10 suggests a four-factor solution, accounting for 71% of the total variance, which is interpretable and makes the best sense.

The first factor is identifiable as innovativeness, the second as change seeking, the third as cognitive effort, and the fourth appears to be a social desirability factor. The four factors are weakly correlated with each other.

Figure 6-3
Scree Test for the Trait Data

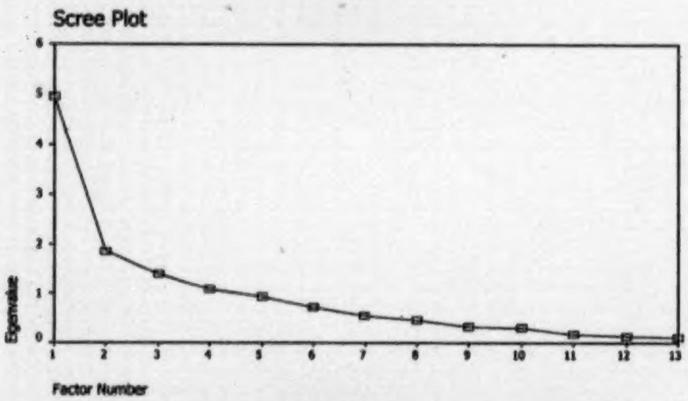


Table 6-10
Rotated Factor Matrix and Alpha of the Trait Measures

Trait scale	Scale Alpha	Innovative -ness	Change seeking	Cognitive effort	Social desirability
Desire for unique consumer products	.86	.678			
Consumer independent judgment making	.88	.667			
Domain-specific innovativeness	.79	.548	.317		.334
Consumer novelty seeking	.90	.537			.607
Buying impulsivity	.87	.461			
Preference for consistency	.84		-.838		
Exploratory acquisition of products	.86		.772		.382
Change seeker index	.88		.678		
Need for precision	.74			.947	
Need to evaluate	.74	.300		.683	
Involvement	.79				.626
Exploratory information seeking	.84				.490
Social desirability	.64				.489
Factor		Correlations among factors			
Innovativeness		1.000	.257	.126	.116
Change seeking			1.000	.369	.291
Cognitive effort				1.000	.311
Social desirability					1.000

Pattern coefficients of below .30 not reported to clarify the factor structure because .30 is the normal cut-point between salient and non-salient pattern coefficients.

1. Traits that load strongly on the innovativeness factor are the Desire for Unique Consumer Products, Consumer Independent Judgment Making, Domain-Specific Innovativeness, Consumer Novelty Seeking, and Buying Impulsivity.
2. Traits that load strongly on the change seeking factor are the Preference For Consistency (which has a negative loading), Exploratory Acquisition of Products, and Change Seeker Index,
3. Traits that load on the cognitive effort factor are the Need For Precision and Need to Evaluate Scale,
4. Traits that load strongly on a social desirability factor are Involvement, Consumer Novelty Seeking, Exploratory Information Seeking and Social Desirability. But there are also secondary loadings for Exploratory Acquisition of Products, and to a less extent Domain-Specific Innovativeness.

Somewhat surprising, Involvement loads strongly only on the social desirability factor. Also surprising is that Consumer Novelty Seeking loads more strongly on the social desirability factor than it does on the innovativeness factor. In addition, several other traits have substantial cross loadings on social desirability. It appears many of these scales are susceptible to social desirability bias, presumably because they include fairly transparent items where people can identify how they would be expected to respond. Note that Buying Impulsivity loads strongly on the innovativeness related factor, even though it is not usually considered to be an innovativeness scale. Buying Impulsivity measures a consumer's tendency to buy spontaneously, unreflectively, immediately, and kinetically. Perhaps, as suggested by Rook and Fisher (1995), highly impulsive buyers are more receptive to the sudden, unexpected buying opportunities.

To investigate whether any traits are associated with reported concept evaluations, I specified a mixed model and used the SPSS mixed effect module to estimate the fixed effect of each trait on concept evaluations. Trait scores are modeled as a covariate, while respondents, concepts, items and occasions are specified as random factors. A separate mixed model is estimated for each trait measure. The mixed model is specified as:

$$y = \beta x + \sum_{j=1}^{4} b_j z_j + \varepsilon$$

where y is the dependent variable (item score or concept score), x is the fixed effect regressor (trait score), β is the fixed effect coefficient and z is the random effect regressor, $j=1,2,3,4$ because there are four random factors in the model, ε is the random error.

Appendix A-8 reports the complete output from the model for Domain-Specific Innovativeness. The fixed effect estimate for Domain-Specific Innovativeness was extracted from the output and reported in Table 6-11 (as shown by the numbers of the first line). The same thing was then done for each of the models involving different traits. Thus Table 6-11 summarizes the estimates for the fixed trait effects from 13 models. If a fixed trait effect is significant and positive, subjects scoring higher on the trait report significantly more favorable concept evaluations. As shown in Table 6-11, there is a significant positive relationship for all the innovativeness-related traits except Consumer Independent Judgment Making and for Involvement.

Table 6-11
Summaries of the Effects of Traits on Concept Evaluations

Parameter	Effect	*S.E.*	*t*	*Sig.*	Lower bound[a]	Upper bound[a]
Innovativeness related:						
Domain-specific innovativeness	.424	.130	3.274	**.002**	.166	.682
Consumer independent judgment making	.087	.123	0.712	.478	-.157	.332
Consumer novelty seeking	.435	.119	3.668	**.000**	.199	.671
Desire for unique consumer products	.359	.145	2.468	**.016**	-.069	.648
Impulsivity	.306	.143	2.133	**.036**	.020	.591
Change seeking related:						
Exploratory acquisition of products	.226	.148	1.525	.131	-.069	.521
Change seeker index	.254	.156	1.633	.107	-.056	.564
Preference for consistency	.193	.139	1.384	.171	-.085	.470
Cognitive effort related:						
Need to evaluate	.279	.238	1.174	.244	-.194	.752
Need for precision	-.065	.224	-.291	.772	-.511	.380
Social Desirability related:						
Social desirability	.276	.207	1.329	.188	-.138	.689
Exploratory information seeking	.291	.148	1.967	.053	-.004	.587
Involvement	.344	.169	2.035	**.045**	.007	.681

[a] Lower and upper bounds for 95% confidence interval for the estimate. Numbers in bold are significant results

Trait Segment Differences in Concept Evaluation Scores

To investigate the effects of screening respondents using the trait measures, I segmented the respondents using a tertiary split of the scores on each trait. Those scoring in the upper third on a trait were classified as high, and those in the lower third were classified as low and the middle third was left out. Table 6-12 shows the mean evaluations reported by these low and high respondent segments for each trait. The results for all concepts show that there is a consistent pattern of higher mean evaluations being reported by the high segments. The trait segment differences in the concept evaluation scores are tested using the two independent sample *t*-test. The differences between the segments are significant for all of the traits except Preference for Consistency and Need for Precision. For major innovations, consumers who are high on the Innovativeness scales and the Change Seeking scales report a significantly higher mean evaluation of concepts. For minor innovations, consumers who are high on Innovativeness scales and Change Seeking scales also report a significantly higher mean evaluation of concepts, except for the Consumer Independent Judgment Making and Preference for Consistency scales.

(Insert Table 6-12 about here)

Therefore, Hypothesis 4a that consumers who are higher on Innovativeness scales report a higher mean evaluation of concepts is supported for all innovativeness related traits, namely DSI, CIJM, CNS, DUCP and Impulsivity in testing of major innovations, but not supported for CIJM in testing of minor innovations. Hypothesis 5 that consumers who are higher on Change Seeking scales report a higher mean evaluation of concepts is supported for all Change Seeking scales, namely EAP, CSI and PFC in testing of major innovations, but not supported for PFC in testing of minor innovations.

Table 6-12 also provides a further breakdown in the trait segment evaluations by concept newness (major versus minor innovations). As shown by comparing the last two columns of Table 6-12, the differences between the high and low segments are greater for major than for minor innovations for all the innovativeness related scales. The direction of the difference is inconsistent for other traits, and in the case of Need For Precision, indicates that there is a significant interaction effect between the trait and concept newness.

Trait Segment Differences in the Generalizability of Concept Testing

When concept tests are used to estimate the concept's market potential or to eliminate poor concepts (Crawford and Di Benedetto 2003), the concepts themselves are the object of measurement. Table 6-13 compares the G coefficients for low and high trait segments when scaling concepts. The high and low trait respondent segments are the same high and low groups that were used in the previous sub-section of this chapter. The specific test design chosen for this comparison is the one hundred respondents who evaluated all the concepts on the six evaluation items on one single occasion. This design is consistent with the number of respondents and items that marketing researchers currently use when conducting concept tests (see Chapter 3).

(Insert Table 6-13 about here)

The results are consistent with the expectation that consumers who are higher on the Innovativeness scales and on the Cognitive Effort scales provide higher quality concept testing data (a higher G coefficient) when testing of all the concepts together or when testing the groups of minor and major innovations separately. The two exceptions are for NFP in the testing of all concepts and CIJM in the testing of minor innovations. The differences in the G coefficients are more extreme for major innovations.

In addition, the trait screening criteria can be ranked for their ability to identify respondents who provide the highest quality concept testing data (G coefficients above 0.90 for the typical design). Respondents who are higher on Domain-Specific Innovativeness provide the best quality data (G coefficient of 0.908) when testing all types of concepts. Respondents who are higher on Buying Impulsivity or Domain-Specific Innovativeness provide the best quality data (G coefficient of 0.921 and 0.911) when testing major innovations. Respondents who are high on Need to Evaluate or Exploratory Acquisition of Products provide the best quality data (G coefficient of 0.900) when testing minor innovations.

The tests of hypotheses about G coefficients are summarized in Table 6-9. There is mixed support for Hypothesis 4b, Hypothesis 6 and Hypothesis 7. Hypothesis 4b that consumers who are higher on Innovativeness scales provide a higher G coefficient for both minor and major innovations, especially for major innovations, is supported for almost all Innovativeness scales in testing of all concepts, testing of minor innovations and testing of major innovations. The only exception is that Hypothesis 4b is not supported for CIJM in testing of minor innovations. Hypothesis 6 that higher cognitive effort segments provide a higher G coefficient is supported for NES in all testings and NFP when testing major innovations and minor innovations separately, but not supported for NFP when testing all types of concepts. Hypothesis 7 that consumers less susceptible to Social Desirability bias provide a higher G coefficient is only supported for SD when testing major innovations, but not supported for SD when testing all concepts or minor innovations. Hypothesis 7 is also not supported for Involvement and EIS in all the testings. Note that here I can only report results of hypotheses about G coefficients without addressing whether the observed difference are statistically significant. Burdick and Graybill (1992) provide a detailed discussion of procedures for establishing confidence intervals on ratios of variance components under normality assumptions. But they started with such simple designs as a $p \times i$ design (see p. 128-129) and didn't have full solutions for generalizability coefficients of any design. For the complicated design I have, the computation is getting more complex and there are currently no suitable statistical tests that can be provided by any advanced statistical software.

Conclusions and managerial implications

This study attempted to examine individual differences in the generalizability of concept testing. The findings led to the following conclusions:

Firstly, I found that there is a significant positive relationship between concept scores and Involvement and all the innovativeness-related traits except Consumer Independent Judgment Making. The fixed effect of each of these traits is significant and positive, indicating that subjects scoring higher on the trait report significantly more favorable concept evaluations.

Second, there is a consistent pattern of higher mean evaluations being reported by the high segments. The differences between the high and low segments are significant for all of the traits except Preference for Consistency and Need for Precision. To be more specific, for major innovations, consumers higher on Innovativeness scales and Change Seeking scales report a significantly higher mean evaluation of concepts. For minor innovations, consumers higher on Innovativeness scales and Change Seeking scales also report a significantly higher mean evaluation of concepts except for Consumer Independent Judgment Making and Preference for Consistency. Moreover, the differences between the high and low segments are greater for major than for minor innovations for all the innovativeness related scales. The direction of the difference is inconsistent for other traits, and in the case of Need for Precision, indicates that there is a significant interaction effect between the trait and concept newness.

Third, the study provides evidence of substantial differences in the generalizability of concept testing for trait-based segments of respondents. In particular, consumers who are higher on Innovativeness scales and on Cognitive Effort related traits provide better quality concept testing data in testing of all concepts except for NFP. This pattern holds when testing minor and major innovations together or separately. The higher trait level segments provide higher G coefficients than the low trait level segments except for CIJM in the testing of minor innovations.

The findings suggest that product managers can identify segments of potential customers who provide higher quality concept testing data. For example, people who are high on Domain-Specific Innovativeness provide the highest quality data when choosing the most positively evaluated concept. People who are high on Buying Impulsivity or Domain-Specific Innovativeness provide the best quality data for concept testing of major innovations, while those who are high on Need to Evaluate Scale or Exploratory Acquisition of Products discriminate best for minor innovations.

Finally, the effects of innovativeness on reported concept evaluations and differences in the generalizability of concept testing for trait based segments are more extreme for major innovations, supporting the claim that subject selection is a more critical issue in concept testing of major innovations.

These findings have implications for subject selection to ensure a concept test provides a required degree of psychometric quality. Here I give several examples to show the value of my work.

1. How many more respondents are needed from a low segment to produce data of the same quality as from a high segment? Take DSI as an example. Selecting 20 respondents who are high on DSI for a concept test of major innovations provides a G coefficient of 0.760 (when 20 respondents evaluate major innovations using 6

evaluation items). If limited to using low DSI respondents for the same concept test, it would take 2000 respondents to evaluate the same major innovations on the same 6 evaluation items to provide data of nearly the same quality (a G coefficient of 0.736).

2. Identify some segments where respondents provide no useful data for a particular decision. As shown in Table 6-13, the low DUCP segment provides garbage (no useful) data when testing major innovations, as the G coefficient is zero.
3. Which design provides higher quality data, sampling 100 respondents (the number of respondents currently used in concept testing practice) to select the best 30 on a scale and then just using the 30 for a concept test of major innovations (design A), or using the original 100 respondents for the test (design B)? Let me assume the screening cost for each respondent is $5 and the testing cost for each participant is $10. Again take DSI as an example. The method to identify the better design is to compare the approximate cost and the G coefficient. First I compare the approximate cost for the two designs. The approximate cost for design A is $800 (100 times $5 plus 30 times $10) while the cost for design B is $1000 (100 times $10). Then I compare the data quality provided by the two designs. Both design A and designs B provide a G coefficient of 0.80. Therefore, design B costs more than design A for data collection but provides the same data quality.

Table 6-12
Trait Segment Differences in Mean Concept Scores

Personality Traits	All concepts		Major innovations		Minor innovations		Differences (High-low)	
	Low	High	Low	High	Low	High	Major	Minor
Innovativeness related:								
- Domain-specific innovativeness	4.644	5.298**	4.644	5.438**	4.645	5.158**	.794	.513
- Consumer independent judgment making	5.047	5.130*	5.015	5.273**	5.079	4.986	.258	-.093
- Consumer novelty seeking	4.663	5.376**	4.685	5.487**	4.640	5.266**	.802	.626
- Desire for unique consumer products	4.787	5.400**	4.850	5.551**	4.724	5.249**	.701	.525
- Impulsivity	4.743	5.220**	4.779	5.289**	4.708	5.151**	.510	.443
Change/consistency seeking related:								
- Exploratory acquisition of products	4.815	5.215**	4.858	5.323**	4.771	5.107**	.465	.336
- Change seeker index	4.654	5.278**	4.734	5.312**	4.575	5.244**	.578	.669
- Preference for consistency	4.897	4.943	4.972	5.079*	4.823	4.812	.107	-.011
Cognitive effort related:								
- Need to evaluate	4.791	5.026**	4.877	5.119**	4.705	4.934**	.242	.229
- Need for precision	4.984	5.006	5.171**	4.965	4.797	5.048**	-.206	.251
Social desirability:								
- Social desirability	4.808	4.932**	4.922	5.013	4.695	4.850**	.091	.155
- Exploratory information seeking	4.782	5.061**	4.872	5.148**	4.692	4.975**	.276	.283
- Involvement	4.765	5.234**	4.813	5.292**	4.716	5.176**	.479	.460

** p<0.01; * p<0.05. All tests are two independent sample *t*-tests (two-tailed). Readers can refer to p. 25 for the test statistic.

Table 6-13
Trait Segment Differences in G-coefficients

Personality Variables	All concepts		Minor innovations		Major innovations	
	Low	High	Low	High	Low	High
Innovativeness related:						
Domain-specific innovativeness	.694	**.908**	.712	.890	.351	**.911**
Consumer independent judgment making	.765	.833	.851	.837	.429	.827
Consumer novelty seeking	.248	.851	.690	.885	.000	.793
Desire for unique consumer products	.144	.872	.679	.733	.000	.858
Impulsivity	.655	.876	.537	.793	.711	**.921**
Change seeking related						
Exploratory acquisition of products	.000	.866	.000	**.900**	.338	.799
Change seeker index	.682	.830	.783	.822	.346	.835
Preference for consistency	.546	.866	.799	.887	.339	.789
Cognitive effort related:						
Need to evaluate	.000	.869	.716	**.900**	.000	.801
Need for precision	.872	.821	.429	.890	.612	.708
Social desirability:						
Social desirability	.858	.868	.845	.894	.862	.823
Exploratory information seeking	.709	.833	.649	.861	.752	.808
Involvement	.616	.842	.784	.816	.000	.784

Numbers in bold are discussed in the text.

6.5 Design Issues Revisited

In this section the design issues addressed in the secondary data studies are re-examined using primary data. The original design issues to be revisited are:

(1) How many respondents are needed to reliably scale concepts?

(2) Is it worth collecting multiple responses rather than relying on a single response item?

(3) Is purchase intent consistently the best response measure to use in concept testing?

The four facets of variation sampled in the primary study are ***concepts***, ***respondents***, ***occasions*** and ***evaluation items***. The concepts evaluated were consumer appliances, selected to include both minor and major innovations. The respondents were IOCS members who signed up for the study. The evaluation items were six seven-point semantic differential scales with both end points labeled asking about liking, feature importance, uniqueness, problem solving ability, believability and purchase intention. From a G theory perspective, the four facets are crossed with each other (Refer to Appendix 9 Figure A-8 for the design represented by the Venn diagram).

I assume the study randomly sampled from a large population of conditions for each of the four facets and use GENOVA (Crick and Brennan 1983) to estimate the variance components for the fifteen estimable main and interaction effects shown in Table 6-14. To aid in interpretation the Table 6-14 also shows the percent of variance due to each source. Table 6-14 is similar to Table 6-5 except that it doesn't include the facet of concept newness. The percent of variance due to concepts is 1.5%, indicating that other effects contribute most of the variance. If the primary managerial purpose is to assess the relative attractiveness of the concepts, concepts become the object of measurement. This means it is desirable for the effect due to concepts to be large and significant. However, only 1.5% of the total variance was due to concepts. To increase discrimination between the concepts in future D studies would require a tremendous sample size of respondents. For example, if 100 respondents evaluate the concepts using six evaluation items on one occasion, the expected G coefficient for concepts is only 0.748. Even 1000 respondents evaluating concepts using six evaluation items on one occasion is only able to reach an expected G coefficient level of 0.857. It seems there is not much gained by sampling more respondents.

The substantial sources of error variance are respondents (account for 21.9% of the total variance), concepts by respondents (18.0%), concepts by respondents by occasions (15%) and residual error confounded with four-way interaction among concepts, respondents, occasions and items (14.4%). The variance due to respondents is large, suggesting that concept scores for respondents differ substantially after averaging over concepts, items and occasions. A large variance component for concepts by respondents indicates that the appeal of a concept varies substantially by segment of respondents. The occasion effect (three-way interaction among respondents, concepts and occasions) is also a substantial contributor to the total variation of concept scores, indicating that respondents evaluate concepts differently on different occasions.

Table 6-14
Variance Estimates and Percent of Variance

Sources of variance	Estimate	Standard error	%
1. Concepts (C)	.058	.032	1.8
2. Occasions (O)	.027	.021	0.8
3. Respondents (R)	.707	.141	21.9
4. Items (I)	.208	.115	6.4
5. CO	.000	.002	0.0
6. CR	.579	.043	18.0
7. CI	.047	.011	1.5
8. OR	.173	.027	5.4
9. OI	.000	.000	0.0
10. RI	.210	.020	6.5
11. COR	.482	.021	15.0
12. COI	.000	.001	0.0
13. CRI	.199	.009	6.2
14. ORI	.071	.006	2.2
15. CORI	.464	.008	14.4
Total	3.867		100.0

Most organizations use the same items in all of their concept tests, making items a fixed rather than a random facet. To determine which particular item performs the best, I also estimate the variance components separately for each evaluation item. As shown by the percent of variance results in Table 6-15, the uniqueness item discriminates best for scaling concepts, as concepts account for 4.8% of its variance. The believability and purchase intent items provide better G coefficients for scaling concepts than the other items. The problem-solving item is the least effective item for scaling concepts. The findings are quite different if the objective is to scale respondents, with believability (42.2%) most effective and liking (29.2%) the least effective item. The G coefficient for concepts when averaging over all six fixed items (0.90) is significantly larger than that when using a single item measure, indicating that averaging over all six items provides a substantial advantage over a single item measure. The expected G coefficient when averaging over six fixed items (0.90) is also greater than that for the six random items (0.75), because the variance due to items by concepts becomes part of universe score variance and it is removed from the definition of the relative error term.

In summary, revisiting the design issues using the primary data strongly confirms the conclusions that were drawn from the secondary data studies:

First, it needs a tremendous number of respondents to reliably scale concepts since the variance due to concepts is generally trivial. Second, averaging over the items provides more reliable information than relying on a single item, because the expected G coefficient when relying on any single item is far lower than if aggregating over six items. Third, what item is best depends on the data and research objective. In the primary data, it appears the believability item performs the best when scaling both concepts and respondents. The popular purchase intention is not the best single item to use.

Table 6-15
Comparisons of Percent of Variance for Different Items

Sources of variance	Liking	Importance	Unique	Solving	Believe	PI	All six items (random)	All six items (fixed)
Concepts (C)	0.9	0.2	4.8	0.0	1.6	0.5	1.8	1.9
Occasions (O)	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.8
Respondents (R)	29.2	34.6	31.5	34.5	42.2	37.3	21.9	21.9
CO	1.4	3.4	6.0	2.7	1.3	1.8	0.0	0.0
CR	4.6	4.7	7.7	3.1	7.1	3.8	18.0	18.1
OR	0.0	0.0	0.0	0.0	0.0	0.0	5.4	5.4
COR	63.9	57.1	50.0	59.7	47.8	56.7	15.0	16.5
I	-	-	-	-	-	-	6.4	6.1
CI	-	-	-	-	-	-	1.5	1.4
OI	-	-	-	-	-	-	0.0	0.0
RI	-	-	-	-	-	-	6.5	6.2
COI	-	-	-	-	-	-	0.0	0.0
CRI	-	-	-	-	-	-	6.2	5.9
ORI	-	-	-	-	-	-	2.2	2.1
CORI	-	-	-	-	-	-	14.4	13.7
G coefficient for concepts[a]	0.30	0.06	0.42	0.01	0.47	0.45	0.75	0.90

[a] For the design of 100 respondents evaluating each concept on one occasion

6.6 Validation Studies

My secondary data studies and primary study focus on the internal psychometric analysis of concept testing, without considering the issue of external validity. A skeptic might argue that such internal analyses could lead product managers to take actions that make the tests internally better, but at the expense of external validity. For example, while the results suggest that a multiple item measure provides greater internal consistency, it might perform worse than the traditional purchase intention item in a predictive validity test.

To address this validation issue I include some validation choice tasks in each wave of the primary study. Ideally, I would have validation data from commercial market introduction as the predictive criterion, but it is very difficult to obtain such data. This is because the time between concept testing and introduction is indefinite and, in any case, only a "biased" fraction of the tested concepts (winners) are ever introduced. There is another complication in external validation. The market performance of the product when introduced is not only a reflection of the concept itself, it also depends on the firm's ability to design and implement an appropriate marketing mix for the new product launch.

In my validation studies, I have three different validation tasks labeled GiftC, GiftR and GiftCM, as described in Table 6-16. Readers may see Appendix 4 for the exact wording of each validation choice task. On the first and second occasion, I use the same GiftC scenario that involves a store gift certificate that provides enough money for respondents to pay for two products among the ten tested products. I include another similar scenario called GiftR on the second occasion that involves a free gift registry service under which respondents can get any of these ten tested products as a gift. It is inevitable that these two scenarios don't simulate the situation when consumers would make a choice with a budget constraint. Practically I can't reveal the price information of the products on the first two occasions to ensure a consistent concept presentation without pricing information across the three occasions. However, at the end of the third occasion after the conventional retest of the concepts, I am able to include a new GiftCM scenario that has a budget constraint as well as the same GiftR scenario as the one on the second occasion. Therefore the validation criterion differences may be due to the explicit identification of a budget constraint. I expect these three different validation tasks can cover all the situations consumers would face when making the choice.

Table 6-16

Description of Validation Choice Tasks

Task	Label	Time	Description of the task
Occasion 1: Task 1	GiftC	First wave	A store gift certificate; Have enough money to pay for two products; Select two products you will actually buy.
Occasion 2: Task 1	GiftC	Second wave	A store gift certificate; Have enough money to pay for two products; Select two products you will actually buy.
Occasion 2: Task 2	GiftR	Second wave	A free gift registry service; Can get any of these ten products as a gift; Select the products you will register.
Occasion 3: Task 1	GiftCM	Third wave	A free gift certificate; Have $1800 financial limit to pay for the products; Select the products you will actually buy
Occasion 3: Task 2	GiftR	Third wave	A free gift registry service; Can get any of these ten products as a gift; Select the products you will register.

In this section, I would like to address the following validation questions:

(1) Does the set of six items (index of six items) predict future choice better than each of the single items?

(2) Does one item predict future choice better than the others?

(3) Does the item providing the best discrimination also have the best predictive validity?

In addition, it would be nice to be able to tell whether the explicit identification of a budget constraint provides more useful validation data. The criteria for a good validation task are that (1) it helps discriminate between the items or individual items and an index, and (2) it does so just as well predictively as concurrently.

Correlation between Mean Concept Scores and Choice Shares

My validity study begins with identifying the aggregate level relationships between the concept evaluation scores (including mean scores and Top 2 Box Scores more commonly employed by practitioners) and the choice shares across the sets of 10 concepts within each occasion (concurrent validity), and using evaluation scores on one occasion to predict validation choice shares on later occasions (predictive validity). The aggregate level Correlational analysis was conducted for the index of evaluation items, and for each evaluation item.

Table 6-17 reports the aggregate level validity (correlation) results for the index of items and for each evaluation item. High positive correlations are indicative of greater validity. Given I am critical of practitioners who use percent Top 2 Box Scores to summarize the rating scale data, it would be interesting to investigate their success in predicting later choice. As shown in the Table 6-17, the uniqueness item provides the best concurrent validity, while the liking item and the set of the items provide the best predictive validity, after averaging over all the validation evidence. Other than the uniqueness item, the set of items provides the better concurrent and predictive validity than most other items when used individually. However, the uniqueness item and the set of the items have poorer concurrent and predictive performance when the budget constrained GiftCM scenario is used. The negative correlations between the uniqueness ratings and the choice shares suggests that consumers are not prone to buy unique products when they have a specific budget limit, presumably because price is a very important consideration when making the purchase decision. For the GiftC and GiftR scenarios, the uniqueness item provides the best concurrent and predictive validity. For the GiftCM scenario, the purchase intent item is the best measure that gives the best concurrent and predictive validity. Since the believability item performs the poorest in both concurrent and predictive validity, I suspect that the set of five items, with the believability item excluded, could be as useful as the set of six items. The results shown in the last column of Table 6-17 confirm my expectations.

Moreover, the results for Top 2 Box Scores and mean concept scores are quite consistent in both concurrent and predictive validity. After averaging all validation evidence, percent Top 2 Box Scores are not better than mean scores in predicting later choice. The exceptions are the percent Top 2 Box Scores of the liking and purchase intention in concurrent validity and the liking in predictive validity. For future work it would be useful to consider what is learned about the validation tasks. First, the GiftCM scenario is quite different from the other tasks, and second GiftR seems interesting as it discriminates between the items just about as well predictively as it does concurrently.

(Insert Table 6-17 about here)

Binomial Logistic Regression between Choice and Concept Scores

Another method to address validation questions is to use logistic regression between choice and items (or the index of items) using individual level data. The validation questions are dichotomous choice responses (yes or no) for which binomial logistic regression is appropriate for use when the dependent variable is a dichotomy and the independent variables are of any type. Logistic regression can be used to predict choice on the basis of continuous concept scores (or six different item scores) and to rank the relative predictive power of the six items.

The dependent variable is choice, which is dichotomous. It can take the value 1 with a probability of chosen *q*, or the value 0 with probability of not chosen 1-*q*. The relationship between the predictor and response variables is not a linear function; instead, the logistic regression function is the logit transformation of *q*:

$\theta = \dfrac{e(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i)}{1 + e(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i)}$ where α is a constant of the equation and $\beta_j, j = 1,2,...i$, are the coefficients of the corresponding predictor variables.

Two logistic regression models are specified to include two different types of predictor variables (the index computed from the set of items and the six different items), respectively:

(1) $\log it[\theta(x)] = \log\left[\dfrac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1$ where x_1 is the index computed from the six items. and

(2) $\log it[\theta(x)] = \log\left[\dfrac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$ where x_1 to x_6 indicate six item scores.

Table 6-17

Concurrent and Predictive Validity at Aggregate Level

	Index (6 items)		Liking		Importance		Unique		Solving		Believe		Purchase		Index (5 items)	
	Mean	T2B	Mean	T2B	Mean	T2B	Mean	T2B	Mean	T2B	Mean	T2B	Mean	T2B	Mean	T2B
Concurrent Validity Tests																
Occasion 1: GiftC	.779	.752	.655	.694	.732	.629	.949	.946	.450	.505	.163	.067	.425	.521	.819	.810
Occasion 2: GiftC	.614	.584	.494	.553	.391	.161	.927	.897	.274	.156	.330	.214	.315	.483	.628	.376
Occasion 2: GiftR	.473	.406	.407	.442	.177	-.070	.862	.830	.066	-.089	.358	.202	.191	.329	.471	.397
Occasion 3: GiftCM	.008	.095	.083	.275	.126	.250	-.548	-.588	.256	.381	.252	.468	.268	.231	-.029	.014
Occasion 3: GiftR	.681	.646	.551	.476	.571	.433	.838	.765	.403	.256	.382	.251	.506	.609	.706	.679
Average	***.511***	***.497***	***.438***	***.488***	***.400***	***.281***	***.606***	***.570***	***.290***	***.242***	***.297***	***.240***	***.341***	***.435***	***.519***	***.455***
Predictive Validity Tests																
Occasion 1-2: GiftC	.608	.575	.500	.563	.640	.509	.880	.851	.269	.316	.016	-.037	.156	.232	.655	.491
Occasion 1-2: GiftR	.452	.404	.378	.398	.427	.290	.828	.809	.036	.036	-.026	-.053	-.016	.039	.489	.463
Occasion 1-3: GiftCM	-.011	-.011	.152	.147	-.034	.079	-.584	-.569	.380	.175	.314	.453	.428	.220	-.051	-.095
Occasion 1-3: GiftR	.782	.702	.694	.682	.620	.487	.865	.828	.451	.423	.381	.240	.534	.579	.793	.724
Occasion 2-3: GiftCM	.105	.073	.305	.205	.277	.381	-.481	-.519	.284	.160	.185	.378	.377	.120	.092	.005
Occasion 2-3: GiftR	.624	.601	.533	.669	.374	.292	.865	.866	.296	.156	.418	.168	.352	.375	.627	.644
Average	***.427***	***.391***	***.427***	***.444***	***.384***	***.340***	***.396***	***.378***	***.286***	***.211***	***.215***	***.192***	***.305***	***.261***	***.434***	***.372***

T2B- Top 2 Box Scores. Numbers in bold are correlations after averaging all validation tasks.

Logistic regression applies maximum likelihood estimation after transforming the choice into a logit variable (the natural log of the odds of the dependent variable occurring or not). In this way, logistic regression estimates the probability of a certain choice occurring.

Table 6-18 summarizes the results from two logistic regression models. Results in bold are for the model where the overall concept score (the index of the six items) provided by each individual are the estimates of the utility for each of the ten concepts. Alternatively the six different items are six different estimates that can be compared for their importance in predicting the choice.

Analysis is run on the standardized data within individuals. The reason for using standardized data is that the G study results in Table 6-15 show that variance due to main effect of respondents is a substantial contributor to the total variation, indicating high heterogeneity among respondents. I standardize the data for each respondent to remove the heterogeneity in their use of the scale.

Similar to the analysis at the aggregate level, predictions can once again be made from occasion 1 to occasions 1, 2 and 3; from occasion 2 to occasions 2 and 3; from occasion 3 to occasion 3. The second column of Table 6-18 reports parameter estimates (*b* coefficients). As shown in the case of O1-O1 (GiftC), the logit for a given choice is 1.801, which means that a unit increase in the concept score is associated with a 1.801 change in the log odds of the choice (the natural log of the probability that choice =1 divided by the probability that the choice = 0). The most common way of interpreting a logit is to convert it to an odds ratio using the function of exp(). When the logit is transformed into an odds ratio, it may be expressed as a percent increase in odds. The odds ratio that corresponds to a logit of 1.801 is 6.054 (e to the 1.801 power), allowing one to say that a unit change in concept scores increases the odds of choice about six times, controlling other variables in the model. Comparing the sixth column in Table 6-18, one can conclude which item or the index of items predicts the best for the choice. I observe that the index computed from the six items has a greater Exp (B) than all the items when used individually, indicating that the index of the items has better concurrent validity and predictive validity than any single item. The liking item has a greater Exp (B) in both concurrent and predictive validity than any other single item, indicating that it has the strongest relationship with consumers' choice. Besides liking, the problem-solving item provides the best concurrent validity, while purchase intention and uniqueness provide the best predictive validity, after averaging over all the validation choice tasks.

The Wald statistic provided in the fourth column is used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero). As shown in Table 6-18, only the logit coefficients for the six item index and for liking are always significant at the 0.05 level. The coefficients for believability are always and the coefficients for importance are almost always not significantly different from zero.

In Table 6-18, I also report the Cox and Snell's R-Square measure. Since for a dichotomous dependent variable, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance, the *R*-squared measure, for logistic regressions is not the actual percent of variance explained. Cox and Snell's *R*-Square measure is not a goodness of fit test but rather a measure of the strength of association. From the last column of Table 6-18, I can see that the association between the concept scores and the choice is around 0.2. The model of O3-O3 (GiftR) has the biggest *R* square (0.195 and 0.230) among concurrent validity tests. Among the six items, the logit coefficients for the liking, uniqueness and believability item are significantly different from zero. And the model of O2-O3 (GiftR) when using the concept data collected on the second occasion to predict the choice task of GiftR on the third occasion provides the best test of predictive validity, which has an *R* square of 0.128 and 0.146. In this model only the liking item has a significant non-zero logit coefficient. At the individual level, again GiftR discriminates between the items just as well predictively as concurrently, compared with the other two validation tasks. However, GiftCM with a budget constraint performs quite differently from the other tasks as more Beta coefficients (indicating the correlation between the concept scores and the item estimates) are negative when GiftCM is used. I have insufficient evidence to determine whether it provides a better validation task, though it is closer to the real world choice.

Table 6-18

Results from Two Binary Logistic Regression Models

	Beta	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**R Square**
Concurrent Validity						
O1-O1 (GiftC)	**1.801**	**.185**	**95.093**	**.000**	**6.054**	**.161**
- Liking	.580	.194	8.933	.003	***1.785***	.165
- Importance	-.041	.216	.036	.850	.960	
- Uniqueness	.290	.137	4.487	.034	1.337	
- Solving	.355	.179	3.955	.047	***1.427***	
- Believability	.223	.180	1.536	.215	1.250	
- Purchase intent	.295	.172	2.946	.086	1.344	
O2-O2 (GiftC)	**1.744**	**.183**	**91.067**	**.000**	**5.718**	**.150**
- Liking	.543	.167	10.530	.001	***1.721***	.155
- Importance	.168	.202	.689	.406	1.183	
- Uniqueness	.161	.124	1.671	.196	1.174	
- Solving	.400	.166	5.790	.016	***1.492***	
- Believability	.145	.182	.634	.426	1.156	
- Purchase intent	.203	.161	1.593	.207	1.225	
O2-O2 (GiftR)	**1.444**	**.221**	**42.574**	**.000**	**4.236**	**.067**
- Liking	.856	.224	14.668	.000	***2.354***	.078
- Importance	-.247	.260	.900	.343	.781	
- Uniqueness	.208	.167	1.560	.212	1.231	
- Solving	.231	.212	1.184	.276	1.260	
- Believability	.244	.237	1.054	.305	***1.276***	
- Purchase intent	.097	.201	.234	.629	1.102	

	Beta	S.E.	Wald	Sig.	Exp(B)	R Square
O3-O3 (GiftCM)	**1.022**	**.116**	**77.293**	**.000**	**2.778**	**.109**
- Liking	.522	.136	14.738	.000	***1.685***	.132
- Importance	.175	.131	1.773	.183	1.191	
- Uniqueness	-.117	.101	1.333	.248	.890	
- Solving	.117	.117	1.004	.316	1.125	
- Believability	-.064	.123	.274	.601	.938	
- Purchase intent	.215	.119	3.277	.070	***1.240***	
O3-O3 (GiftR)	**1.645**	**.149**	**122.082**	**.000**	**5.183**	**.195**
- Liking	.795	.164	23.381	.000	***2.214***	.230
- Importance	.373	.163	5.256	.022	***1.453***	
- Uniqueness	-.187	.115	2.638	.104	.830	
- Solving	.137	.131	1.085	.298	1.147	
- Believability	-.068	.146	.219	.639	.934	
- Purchase intent	.386	.133	8.457	.004	***1.471***	
Predictive Validity						
O1-O2 (GiftC)	**1.224**	**.156**	**61.920**	**.000**	**3.400**	**.093**
- Liking	.373	.168	4.910	.027	***1.452***	.100
- Importance	.026	.197	.017	.896	1.026	
- Uniqueness	.310	.130	5.718	.017	***1.364***	
- Solving	.277	.167	2.748	.097	1.319	
- Believability	-.105	.161	.424	.515	.900	
- Purchase intent	.196	.160	1.499	.221	1.216	
O1-O2 (GiftR)	**.731**	**.183**	**15.904**	**.000**	**2.077**	**.022**
- Liking	.500	.218	5.280	.022	***1.649***	.039
- Importance	-.127	.242	.274	.600	.881	
- Uniqueness	.585	.179	10.643	.001	***1.795***	
- Solving	-.059	.208	.080	.777	.943	
- Believability	-.199	.200	.996	.318	.819	
- Purchase intent	.040	.203	.039	.843	1.041	

	Beta	S.E.	Wald	Sig.	Exp(B)	R Square
O1-O3 (GiftCM)	**.475**	**.104**	**20.782**	**.000**	**1.609**	**.027**
- Liking	.342	.118	8.360	.004	***1.407***	.052
- Importance	-.011	.138	.006	.938	.989	
- Uniqueness	-.153	.091	2.810	.094	.858	
- Solving	.387	.126	9.518	.002	***1.473***	
- Believability	-.098	.116	.712	.399	.906	
- Purchase intent	-.151	.124	1.478	.224	.860	
O1-O3 (GiftR)	**.723**	**.117**	**38.170**	**.000**	**2.060**	**.052**
- Liking	.252	.129	3.830	.050	***1.287***	.065
- Importance	-.045	.151	.088	.767	.956	
- Uniqueness	-.100	.098	1.036	.309	.905	
- Solving	.194	.136	2.032	.154	1.214	
- Believability	.043	.127	.114	.736	1.044	
- Purchase intent	.248	.134	3.398	.065	***1.281***	
O2-O3 (GiftCM)	**.988**	**.119**	**68.460**	**.000**	**2.685**	**.097**
- Liking	.407	.121	11.356	.001	***1.502***	.115
- Importance	.275	.142	3.760	.053	1.317	
- Uniqueness	-.078	.092	.727	.394	.925	
- Solving	.011	.127	.007	.931	1.011	
- Believability	-.053	.129	.169	.681	.948	
- Purchase intent	.297	.126	5.550	.018	***1.346***	
O2-O3 (GiftR)	**1.277**	**.138**	**85.952**	**.000**	**3.585**	**.128**
- Liking	.622	.138	23.080	.000	***1.939***	.146
- Importance	-.006	.158	.001	.972	.994	
- Uniqueness	-.031	.099	.095	.758	.970	
- Solving	.223	.137	2.652	.103	***1.250***	
- Believability	.203	.146	1.946	.163	1.225	
- Purchase intent	.130	.135	.928	.335	1.139	

Numbers in bold are results for the index of six items.

Conclusions

Validation studies at both aggregate and individual levels show that the index of six items predicts better than any items when used individually. At an individual respondent level, the liking item appears to be the best of the six items to use for concurrent and predictive validity. Other relatively effective items are uniqueness and purchase intent for predictive validity and problem solving for concurrent validity. At an aggregate market level, uniqueness is the best item for predictive and concurrent validity.

Of the three validation tasks, GiftCM provides the most distinct criterion. GiftR seems interesting as it discriminates between the items just as well predictively as concurrently. Whether the explicit identification of a budget constraint helps provide more valid data is a question that needs more future work.

In the secondary data studies (Chapter 5), I concluded that the averaging over a set of items provides more generalizable information than relying on a single item. In this section I find that the index of six items predicts better than any other single item. It is not surprising that a summary of the multiple predictors performs better than a single predictor. Usually it is not costly to collect additional items to get more information. Therefore it is wise to collect multiple item measures in a concept test and analyze the concept data on the basis of an index computed from the items rather than relying on one single item.

In Chapter 5, I also concluded that the best single item for discrimination between concepts is dependent on the research situation, but it is not the commonly used purchase intent item. In this section, I find that the liking item and the uniqueness item perform the best for both predictive and concurrent validity at both the aggregate and individual level. Besides these two items, the other effective items are purchase intent item (at the aggregate level) and problem solving (at the individual level) for concurrent validity and importance (at the aggregate level) and purchase intent (at the individual level) for predictive validity. Overall, percent Top 2 Box Scores that practitioners currently use perform no better than mean concept scores for both concurrent validity and predictive validity. The exceptions are the percent Top 2 Box Scores for the liking and purchase intention items in concurrent validity and the liking item in predictive validity.

Chapter 7 Contributions and Future Work

7.1 Primary Contributions and Conclusions

Primary Contributions

The importance of concept testing has been widely recognized by both new product development practitioners and marketing researchers. My dissertation contributes both to the field of new product development and to the field of measurement in marketing. It leverages the synergy between these two fields by using a measurement theory, namely G theory, to re-examine measurement issues in new product concept testing from a measurement and decision making perspective.

When confronting measurement issues, marketing researchers still often rely on the validity of measurement procedures borrowed from psychology even when they are not entirely applicable. For example, psychologists are primarily interested in valid measurement of one or more latent characteristics of individuals. In marketing research, the purpose of measurement is often more complex, and scaling consumers is only one of many things managers need to do. Equally relevant is identifying which concept will obtain the most favorable evaluation from consumers or which customer segment will respond most favorably to a particular concept. In new product concept testing, there are typically multiple objectives, such as to develop further the original idea, to estimate the concept's market potential, to eliminate poor concepts and to identify the highest potential segments. I use G theory as my major methodology to reexamine measurement issues in concept testing because G theory explicitly recognizes the fact that measurement can take place for multiple objects of measurement and over multiple facets of generalizations. Using G theory, multiple purposes of measurement can be specifically accommodated. Besides G theory, other methods such as Correlational and logistic analyses are also used to investigate evidence of validity in validation studies.

In the field of new product development, measurement issues have not been thoroughly addressed. Traditional measurement methods and criteria (e.g., percent Top 2 Box Scores) are still viewed as a panacea, but there is no substitute for establishing the reliability and construct validity of new product development research measures. My dissertation research treats concept testing as a measurement and decision-making process rather than a market research technique. From this new perspective, the purpose of concept testing is generalizing, with a known level of reliability and validity, from a planned set of customer test responses to a defined universe of conditions under which the product could be marketed once developed.

My dissertation consists of several research elements, including

(1) A survey of new product managers and research consultants on the contemporary concept testing practice.

(2) An analysis of secondary concept testing data.

(3) A primary study collecting on-line concept testing data from the same respondents on multiple occasions.

My research builds on previous and ongoing work on new product concept testing and measurement. But rather than simply looking at the new solutions to the measurement issues in isolation, my approach differs from previous work in this area by developing an integrated conceptual framework from a G theory perspective for categorizing the sources of variance that influence the observed scores in concept testing and identify four types of such factors, namely concept related factors, response task factors, situational factors and respondent factors. Then current testing practice can be viewed as making implicit assumptions about the relative size of the various sources of variance.

The survey of new product managers and research consultants is all about describing the current state of concept testing practice. Although it could not fully answer all of the questions about measurement issues in concept testing, the findings of the study do help clarify the current state of concept testing practice. Moreover, it helps me identify the design issues most in need of follow-up research, such as what distinguishes the better tests (the number of respondents needed to reliably scale the concepts, the number of items to use, the best item to use, and the types of respondents who can provide better quality data) and how those relate to consumer trial (generalizability and predictive validity of the test results) – both of which are central for practice. So I conduct the secondary data analyses and the primary study to answer these questions of interest.

First, I examine the contributions that the estimable sources make to overall concept testing variance in my secondary data studies. The results provide guidelines for improving the psychometric quality of concept testing and answer the questions of practitioners about (1) how many respondents are needed to reliably scale the concepts; (2) is it worth collecting multiple responses rather than relying on a single response measure; (3) is purchase intent consistently the best response measure to use in concept testing; (4) what is gained by sampling respondents from multiple locations. However, the secondary data studies fall short of the fact that some sources of variance could not be investigated with any of the secondary data sets. A potentially important source that is not investigated is testing occasions. Product managers assume that consumer evaluations of concepts are generalizable from the time (and research environment) of concept testing to the time (and market environment) of market introduction. But no evidence on the issue was available from the current concept testing literature, as summarized in Table 2-1, nor from the secondary data, as summarized in Table 5-2. Therefore, investigating the generalizability of concept testing results over occasions is a major contribution of my primary research. Another issue not addressable with secondary data is respondent selection. It was unclear whether some types of respondents provided higher quality evaluation data than others for some types of new concepts. My primary study also contributes to the understanding of individual differences in the generalizability of concept testing.

Second, to deal with these fundamentals that were not addressable with secondary data, it was necessary to conduct primary studies to concept test minor and major innovations, over multiple test occasions, using respondents who can be clustered into segments on characteristics assumed to influence their test responses or their predictive capabilities. Moreover, my secondary data studies focused on the internal validity of concept testing without considering the issues of external validity. I had no idea whether which item or the index computed from a set of items that has the best internal validity (e.g., discriminates best among the tested concepts) also has the best predictive validity. Then it was of interest to collect validation data for external criteria in the primary study.

My primary study examines the generalizability of testing over occasions and the individual differences in the generalizability of concept scores, which have seldom been investigated in the concept testing literature. This research provides insight into how far practitioners can rely on a concept test. The answer to this question is important because a time delay between concept testing and market introduction is an inherent feature of new product development and practitioners inevitably need to generalize the customer evaluation of concepts, observed on one occasion, to a whole set of possible occasions (e.g., potential market introduction). If the concept testing results cannot generalize over testing occasions sufficiently well to enable managerial conclusions to be drawn from a single occasion test, the worth of current concept testing practice that relies on a single occasion test is in question. This research also investigates a number of personality traits in the areas of innovativeness, change seeking and cognitive effort that could influence concept evaluation scores and help identify respondents who provide substantially higher quality data in concept testing. The findings provide evidence of individual differences for both minor and major innovations in terms of the mean evaluation of concepts and the quality of concept testing data (G coefficients). The latter has implications for subject selection to ensure a concept test design provides a required degree of psychometric quality.

Besides the theoretical contributions my thesis has made in the field of new product development and the field of measurement in marketing, it is important to appreciate its contributions in methodology. From a methodology point of view, my dissertation is a new marketing application of generalizability theory. It demonstrates the value of a new measurement approach to assessing the psychometric quality of concept testing data. The new criteria in the approach are the variance components obtained in a G study for the facets which can be used to determine the contribution of each facet to the observed variation in concept testing and the G coefficient for different objects of measurement. An issue not fully resolved in G theory is constructing the confidence intervals on variance components and G coefficients. Burdick and Graybill (1992) provide the general equations for calculating confidence intervals on variance components, where variance components are expressed as the linear combinations and/or ratios of expected mean squares. However, it is very complicated to calculate the confidence intervals for multi facet G study designs. Current statistical packages such as SPSS and SAS have not provided such a function for calculating confidence intervals on variance components. I have made it doable by designing a MatLab program (See Appendix 6 for the description and detailed calculation code). To execute the program, readers are simply required to follow the Burdick and Graybill (1992) equations and input the expected mean squares and the corresponding degree of freedom involved in the variance component expression. Now the problem I have solved is the numerical estimation of confidence intervals on linear combinations of expected mean squares. A problem that remains to be unresolved is the estimation of confidence intervals on G coefficients, which involves the linear combination of expected mean squares and ratio of expected mean squares, making the calculation even more complicated. No suitable equations are currently provided in the variance component analysis and G theory literature.

My dissertation has important practical implications for the design and implementation of concept testing and new product development. Furthermore it would make an important contribution to improving the efficiency of new product development. New product development practitioners can learn from the study some ways to improve the psychometric quality of concept testing. Some lessons learned about the design, analysis and interpretation of concept tests may also apply to other customer research situations, such as prototype/use testing, package testing, and promotion testing.

The main conclusions for each part of the thesis in more detail are summarized below.

Main Conclusions

Learning from the survey on contemporary concept testing practice

The survey of new product managers and research consultants sheds light on the state of contemporary practice related to the use of various models and methods of concept testing, their problems and desired improvements, and the evidence for reliability and validity. A typical concept test (1) has multiple objectives, among which to develop further the original idea is primary; (2) usually tests incrementally new product concepts; (3) is equally likely to be a monadic or comparative test; (4) usually includes pricing information in the tested concepts; (5) presents the concept stimuli as stripped descriptions; (6) uses about 92 respondents for each concept; (7) uses both unstructured and structured measures; (8) employs scales with both numerical categories and verbal labels; (9) selects respondents by product class usage, specific product usage or lead user criteria, (10) collects data by face-to-face interview, and (11) assesses the outcome using percent Top 2 Box Scores and/or rating scale mean compared with a company specific norm.

One focus of the survey is to investigate any difference in the approach taken for incremental versus radical new concepts and statistically significant evidence has been found. It looks like concept testings of radical new products require (1) more carefully screening respondents on criteria beyond simple product use or demographics; (2) more commonly used lead user criteria; (3) larger samples of respondents; (4) collection of more detailed evaluation information from each respondent; and (5) more advanced analysis methods. In particular, an average of 195 respondents per concept was reported in radically new product concept tests, five times the number of respondents reported in incrementally new product concept tests. There is more use of both unstructured liking and structured purchase intent and disliking measures for radical concept tests. There is more use of company norms to judge the outcome for radical concept tests, but otherwise the analysis methods are not more appropriate.

It is surprising that product managers are no more satisfied with the predictive performance of their current concept testing approach when testing incremental concepts than when testing radical concepts. Other exploratory findings suggest respondents are more satisfied with concept tests when using more than four structured questions and when using more respondents (greater than 100) to evaluate each concepts.

Most practitioners prefer to keep the evidence for reliability and validity confidential, so there is little public information on concept test performance. This makes primary academic research into the measurement issues in concept testing more necessary and valuable, and raises questions about the transparency of the testing process.

Learning from secondary data studies

Generalizability analysis of four secondary datasets provides new insights into the design of concept tests and the psychometric quality of the concept testing data. As summarized in Table 5-2, I found that (1) the concepts facet is not a major contributor to response variation; (2) of the response task factors, concept formulations are a trivial source of variance, but items are not always a trivial source of variance; (3) the situational factors that are investigated are trivial sources of variance; (4) respondents are always a major contributor to the total variation; (5) concepts by respondents are not always a major contributor and the other interactions are often not trivial and (6) residual error is always a major source of variance.

The analyses of the secondary datasets also enable some useful conclusions to be drawn about the four managerial design questions in concept testing:

First, the results provide evidence that the sample size needed to reliably scale concepts depends on the types of concepts being tested. Usually, the bigger the sample size the more reliable the data. It appears that fewer respondents are needed when testing heterogeneous concepts while concept testing of similar concepts requires more respondents. The design provides no useful data for scaling concepts if the variance component for concepts is zero.

Second, averaging over items provides considerably more reliable information than relying on a single item. An expected G coefficient when aggregating over all items is far higher than when relying on a single item. If the same items are always used as a fixed scale, the G coefficient is even higher. Thus, it would seem unwise to rely on a single evaluation item when comparing concepts.

Third, which specific item is best for that situation is very inconsistent and very context-specific. It depends on the data and research objective. The popular purchase intention is never the best single item to use.

Fourth, not much is gained by sampling levels of the response task factor of concept formulations in the academic study, or levels of the situational factors, such as cities and situational segments, able to be investigated in the industry studies.

Finally, concept testing should be designed to meet the needs of specific managerial tasks. A concept test design that is good for identifying these respondents who are more interested in all of the new concepts may not be good for identifying which concept is better than the rest.

Learning from primary studies

1. The generalizability of testing over occasions

A striking observation in this study is that there is a systematic decline in mean evaluation of concepts from occasion one to occasions two and three. However the fading effect is very weak for the second and the third occasion. Repeated measures analysis of the concept scores indicate that the main effect of occasions is significant but there is no significant interaction effect between occasion and newness. However, the three-way interaction among subjects, concepts and occasions is a substantial contributor to variation in concept testing of both major and minor innovations, indicating that the three-way interaction among subjects, concepts and occasions is a major source of error variability in concept testing. Moreover, the subjects by concepts by occasions effect is more substantial for major innovations than for minor innovations. This large interaction effect indicates that consumers evaluate the same concepts differently on different occasions, and this could be a threat to the validity of concept testing.

Explicitly recognizing occasions as a facet of error variance influenced the generalizability of the test results. I confirm that the data collected on a single occasion leads to overconfidence in the quality of concept testing results. However, the extent to which it alters the psychometric quality of the concept test data varies with the nature of the managerial task (e.g., assessing concepts or segmenting customers). To be specific, consumer evaluation of concepts, observed on one occasion in this study is not generalizable, to a whole set of possible occasions (e.g., potential market introduction) if the purpose of the concept testing is to identify the types of respondents who are most interested in particular concepts (segmentation). However, if the purpose of concept testing is to assess the relative attractiveness of the concepts (scaling concepts), the generalizability from data collected on a single occasion would be acceptable.

To increase the dependability of this generalization, I suggest that managers should conduct at least small scale G studies that sample occasions as a facet of measurement in order to determine the extent to which scores obtained on a single occasion are as generalizable as expected to scores that might be obtained on different, but similar occasions.

2. Individual differences in the generalizability of concept testing

The investigation of individual differences in the generalizability of concept testing found that there is a significant positive relationship between concept scores and Involvement and all the innovativeness-related traits except Consumer Independent Judgment Making, indicating that subjects scoring higher on innovativeness scales and Involvement report significantly more favorable concept evaluations. Another observation is the consistent pattern of higher mean evaluations being reported by the high segments. The differences between the high and low segments are significant for all of the traits except Preference For Consistency and Need For Precision. Moreover, the differences between the high and low segments are more substantial for major than for minor innovations for all innovativeness scales.

My study provides evidence of substantial differences in the generalizability of concept testing for trait-based segments of respondents. The effects of innovativeness on reported concept evaluations and differences in the generalizability of concept testing for trait based segments are more extreme for major innovations. In particular, consumers who are higher on innovativeness and on cognitive effort scales provide better quality concept testing data for both minor and major innovations.

The findings have implications for subject selection to ensure a concept test provides a required degree of psychometric quality. The findings suggest that people higher on Domain-Specific Innovativeness provide the highest quality data when scaling concepts. People higher on Buying Impulsivity or Domain-Specific Innovativeness provide the best quality data for concept testing of major innovations, while those higher on Need to Evaluate Scale or Exploratory Acquisition of Products discriminate best for minor innovations. Other implications include that some low segments provide no useful data for the decision, as the G coefficient is zero. To achieve the same quality of the data as provided by using the high segment respondents, using the low segment respondents might require a tremendous number of respondents. The design of sampling 100 respondents to select the best 30 on a scale and then just using the 30 for a concept test of major innovations might provide the same quality of data with comparatively lower cost than the design of using the original 100 respondents for the test.

3. Design issues revisited

In the design issues revisited section, I confirm the conclusions drawn from the secondary data studies: (1) When the variance due to concepts is trivial, it needs a tremendous number of respondents to reliably scale concepts because there is not much gained by sampling more respondents. (2) Averaging over the items provides more reliable information than relying on a single item measure because the G coefficient when relying on any single item is lower than when relying on the index computed from the items. (3) The purchase intent item is not the best single item to use. In my primary study, the believability item performs the best for scaling concepts and scaling respondents. This result is inconsistent with the secondary data studies about which is the best item. Again it confirms that what item is best is very context-specific and depending on the data and research objectives.

4. Validation studies

I include validation choice tasks in each wave of the primary study. I have three different validation tasks labeled GiftC, GiftR and GiftCM. In GiftC, respondents assume they have a store certificate that has enough money for them to buy two of the ten tested products. In GiftR, respondents assume that they are provided a free gift registry service then they can pick any of the tested products as gifts. In GiftCM, respondents imagine they have a store certificate valued 1800 dollars and can buy any of the tested products. The validity criterion differences may be due to the explicit identification of a budget constraint. I would like to address the validation questions such as (1) Does the index computed from six items predict future choice better than all of the single items? (2) Does one item predict future choice better than the others? (3) Does the item providing the best discrimination also have the best predictive validity?

Validation studies at both aggregate and individual levels show that the index of the items predicts better than any of the items when used on its own. At the individual level, liking is the best item to use for concurrent and predictive validity. Other effective items are uniqueness and purchase intent for predictive validity and problem solving for concurrent validity. At the aggregate level, uniqueness is the best item for predictive and concurrent validity. GiftR seems to discriminate the best between the items or individual items and an index just as well predictively as concurrently. GiftCM with a budget constraint performs quite differently from the other tasks. I have insufficient evidence to determine whether it provides a better validation task, though it is closer to the real world choice decision.

While the best single item for discrimination is dependent on the research situation, the liking item and the uniqueness item are the best to use for both predictive and concurrent validity. Another interesting finding is that percent Top 2 Box Scores that practitioners currently use perform no better than mean concept scores in predicting later choice.

7.2 Future Work

My dissertation research suggests some ways to improve the psychometric quality of concept testing. But, as is usual in the early development of a field (since measurement issues have been neglected in this area), much work remains to be done.

New product managers provided some suggestions to improve concept testing. One of the common themes is to make more use of virtual testing. A related direction to further research is to examine whether the adoption of new techniques in new product evaluation testing can enhance the validity and generalizability of test results. In recent years, virtual environment techniques have developed sufficiently to generate a more realistic test environment and interaction techniques have improved enough to provide natural and intuitive modes of interaction between a test person and its surrounding elements (Dahan and Hauser 2002). An interesting topic is measuring consumer preferences in the context of competitive sets and in more natural shopping environments. Researchers may be able to use virtual environment techniques to simulate particular aspects of the market entry environment, such as early market share, type of word-of-mouth being received, and channel availability, at the time of concept testing. Furthermore, respondents could be provided with the possibility to interact with the test environment. For example, respondents could be able to "go" to different stores, "grab" products from the simulated shelves, to put them back and grab new ones, to put the products into the shopping basket, or to review the comments from the previous users. Then the utility of manipulating aspects of the market entry environment can be assessed at the time of concept testing.

Additional research is needed to more fully address the predictive validity of concept testing. My work is mainly focused on the internal psychometric analyses of concept testing. My validation data are surrogate validation data (Ideally, I would have validation data from commercial market introduction as the predictive criterion), though they do offer a criterion for evaluating the performance of the items. I propose some future work to combine both internal and external criteria to examine the worth of concept test methods. For example, contrasting different concept test methods against how the products actually do in the market place, or even in subsequent product placement tests. Internal analyses are focused on a measurement data set and help understand how that data relates to itself, while external validation analyses enable the testing results to generalize to the population at large. Concept testing comprises two types of errors (Crawford and Di Benedetto 2003). The first type of error occurs when the test results suggest going forward with the tested product when in fact a no-go decision is appropriate. The second type of error occurs when concept-test leads to a no-go decision when in fact a go decision is appropriate. New product practitioners are more interested in how concept tests predict the product success, in which both types of the above errors are avoided. However, the ability of concept testing to avoid the two types of errors is the function of the soundness of the testing design in which internal validity and external validity are equally important. I have proved the usefulness of psychometric properties when applied to an internal criterion. Much empirical research is needed to prove the usefulness of the psychometric properties when applied to an external criterion. Another interesting question in validation studies that needs more future work is whether the explicit identification of a budget constraint helps provide more valid data.

Another issue of interest is building a model to facilitate effective tradeoff decision-making in the design of concept test. Since the G-coefficient for an object of measurement (usually concepts) can be increased by more sampling of conditions of any facets that are sources of error variance, more than one design can produce data of the desired quality. Thus practitioners face tradeoffs in the design of concept test such as whether to sample more respondents, to use more items or to employ the test on multiple occasions. They also face the tradeoff of whether to screen or not to screen respondents. This suggests a model could be built to aid management in finding the best test design, combining estimated variance components with estimates for the cost of sampling for each facet of generalization.

New product development practitioners now have even more exciting challenges and opportunities when facing more face fierce competition, market and product globalization, and more demanding and sophisticated consumers. Current concept testing practice, which still relies on Classical Test Theory based measurement approach, cries for a new perspective on the measurement issues. The understanding and the acceptance of the new measurement approach with which few managers may be familiar is a long process. The critical question for academics is to help practitioners understand what relative advantage these newer approaches have over current, more conventional approaches to concept testing. Demonstration of a compelling competitive advantage to the firms requires a great deal of devotion, work and time.

Finally, I propose a multivariate G approach that can provide covariance terms among the levels of each fixed facet in my follow-up research. The dimensionality shown by the items could be of substantive interest, so it is an interesting topic to be done sometime. But at this stage, univariate G approach has given sufficient results that can solve the managerial problems I raised.

Bibliography

Ainslie, George and Nick Haslam (1992), "Self-Control," in *Choice Over Time,* George Loewenstein and Jon Elster, eds. New York: Russell Sage Foundation, 177-209.

Alexander, David, John G. Lynch, Jr., and Qing Wang (2006), "Temporal Stability in Consumers' Acquisition Intentions for Really New Products," Working paper.

Algina, James and H. J. Kesselman (1997), " Detecting repeated measures effects with univariate and multivariate statistics," *Psychological Methods*, 2, 208-218.

Anderson, James C. (1987), "The Effect of Type of Representation on Judgment of New Product Acceptance," *Industrial Marketing and Purchasing,* 2, 29-46.

Anschuetz, Ned F. (1996), "Evaluating Ideas and Concepts for New Consumer Products," in *The PDMA Handbook of New Product Development,* Milton D Rosenau et al., eds. New York: John Wiley & Sons, 195-215.

Armstrong, Scott J. and Terry Overton (1971), "Brief versus Comprehensive Descriptions in Measuring Intentions to Purchase," *Journal of Marketing Research,* 8 (1), 114-117.

Batsell, Richard R. and Yoram Wind (1980), "Product Testing: Current Methods and Needed Developments," *Journal of the Marketing Research Society,* 22 (2), 115-139.

Baumgartner, Hans (2002), "Toward a Personology of the Consumer," *Journal of Consumer Research,* 29 (September), 286-292.

Baumgartner, Hans and Jan-Benedict E. M. Steenkamp (1996), "Exploratory Consumer Buying Behavior: Conceptualization and Measurement," *International Journal of Research in Marketing,* 13 (2), 121-137.

Bell, Marie (1994), "Nestle Refrigerated Foods: Contadina Pasta & Pizza (A)," in *Harvard Business School Cases,* Boston: Harvard Business School Press.

Bengston, Roger and Henry Brenner (1964), "Product Test Results Using Three Different Methodologies," *Journal of Marketing Research,* 1 (November), 49-52.

Blythe, Jim (1999), "Innovativeness and Newness in High Tech Consumer Durables," *Journal of Product and Brand Management,* 8 (5), 415-429.

Booz Allen Sales Estimating System (BASES) (2005), "PreBASES: Volumetric Early Concept Screening that Works," available at Http://www.bases.com/services/prebases.html.

Brennan, Robert L. (2000), "Performance Assessments from the Perspective of Generalizability Theory," *Applied Psychological Measurement,* 24 (4), 339-353.

---- (2001a), *Generalizability Theory,* New York: Springer-Verlag.

---- (2001b), *Manual for urGENOVA,* Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brown, Tom J., John C. Mowen, Todd D. Donavan and Jane W. Licata (2002), "The Customer Orientation of Service Workers: Personality Trait Effects on Self and Supervisor Performance Ratings," *Journal of Marketing Research,* 39 (February), 110-119.

Burdick, Richard K. and Franklin A. Graybill (1992), *Confidence Intervals on Variance Components,* Marcel Dekker, Inc.

Calantone, Roger J., Kwong Chan and Anna S. Cui (2006), "Decomposing Product Innovativeness and Its Effect on New Product Success," *Journal of Product Innovation Management*, 23 (4), 408-421.

Carvalho, A. (2006), "Diffusion of Mobile Phones in Portugal: Unexpected Success?" *INNOVATION PRESSURE International ProACT Conference 15-17th March 2006*, Tampere, Finland.

Cattell, R. B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research*, 1, 245-276.

Cialdini, Robert B., Melanie R. Trost and Jason T. Newsom (1995), "Preference for Consistency: The Development of a Valid Measure and The Discovery of Surprising Behavioral Implications," *Journal of Personality and Social Psychology,* 69, 318-328.

Cooper, Lee G. (2000), "Strategic Marketing Planning for Radically New Products," *Journal of Marketing,* 64, 1-16.

Corrado, C., W. Dunn and M. Otoo (2006), *Incentives and Prices for Motor Vehicles: What Has Been Happening in Recent Years?* Finance and Economics Discussion Series, Federal Reserve Board, Washington, D. C.

Crawford, Merle C. and Anthony C. Di Benedetto (2003), *New Products Management,* 7th Edition. New York: McGraw-Hill.

Creusen, Maríëlle E. H. and Jan P. L. Schoormans (2005), "The Different Roles of Product Appearance in Consumer Choice," *Journal of Product Innovation Management,* 22, 63-81.

Crick, Joe E. and Robert L. Brennan (1983), *Manual for GENOVA: A Generalized Analysis of Variance System* (ACT Technical Bulletin No. 43), Iowa City, IA: ACT. Inc.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles,* New York: John Wiley & Sons.

Cronbach, Lee J., Nageswari Rajaratnam, and Goldine C. Gleser (1963), "Theory of Generalizability: A Liberalization of Reliability Theory," *British Journal of Statistical Psychology,* 16 (November), 137-163.

Crowne, Douglas P. and David Marlowe (1960), "A New Scale for Social Desirability Independent of Psychopathology," *Journal of Consulting Psychology,* 24, 349-354.

Dahan, Ely and Haim Mendelson (2001), "An Extreme-Value Model of Concept Testing," *Management Science,* 47 (January), 102-116.

Dahan, Ely and John R. Hauser (2002), "The Virtual Consumer," *Journal of Product Innovation Management,* 19 (5), 332-353.

Dahan, Ely and V. Srinivasan (2000), "The Predictive Power of Internet-based Product Concept Testing Using Visual Depiction and Animation," *Journal of Product Innovation Management,* 17 (2), 99-109.

Dahl, Darren W. and Steve Hoeffler (2004), "Visualizing the Self: Exploring the Potential Benefits and Drawbacks for New Product Evaluation," *Journal of Product Innovation Management,* 21 (4), 259-267.

Dickinson, John R. and Carolyn P. Wilby (1997), "Concept Testing with and without Product Trial," *Journal of Product Innovation Management,* 14 (2), 117-125.

Domzal, Teresa J. and Lynette S. Unger (1985), "Judgments of Verbal versus Pictorial Presentations of a Product with Functional and Aesthetic Features," *Advances in Consumer Research,* 12 (1), 268-272.

Duke, Charles R. (1994), "Understanding Customer Abilities in Product Concept Tests," *Journal of Product and Brand Management,* 3 (1), 48-57.

Finn, Adam (1985), "A Theory of The Consumer Evaluation Process For New Product Concepts," *Research in Consumer Behavior,* 1, 35-65.

---- (2006), "Doing A Double Take: Accounting for Occasions in Service Performance Assessment," Working paper.

Finn, Adam and Ujwal Kayande (1997), "Reliability Assessment and Optimization of Marketing Measurement," *Journal of Marketing Research,* 34 (May), 262-275.

---- and ---- (2002), "New Product Concept Testing: A Generalizability Theory Based Perspective," Working Paper.

Fischoff, Baruch (1991), "Value Elicitation: Is There Anything in There?" *American Psychologist,* 46 (8), 835-846.

Garcia, Rosanna and Calantone, Roger J. (2002), "A Critical Look at Technological Innovation Typology and Innovativeness Terminology: A Literature Review," *Journal of Product Innovation Management*, 19 (2), 110-131.

Gardner, P. L. (1975), "Scales and statistics," *Review of Educational Research,* 45, 43-57.

Golder, Peter N. and Gerard J. Tellis (1993), "Pioneering Advantage: Marketing Fact or Marketing Legend," *Journal of Marketing Research,* 30 (May), 158-170.

Goldsmith, Ronald and Charles F. Hofacker (1991), "Measuring Consumer Innovativeness," *Journal of Academy of Marketing Science,* 19 (3), 209-222.

Gourville, John T. (2005), "The Curse of Innovation: A Theory of Why Innovative New Products Fail in the Marketplace," *Harvard Business School Marketing Research Papers* No. 05-06.

Graham, M. B. W. (1986), *RCA & the VideoDisc: The Business of Research*, Cambridge University Press.

Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," Journal of Marketing, 54, 3-19.

Green, Paul E. and Wayne S. DeSarbo (1979), "Componential Segmentation in the Analysis of Consumer Trade-offs," *Journal of Marketing,* 43, 83-91.

Green, Paul E. and Yoram Wind (1975), "New Ways to Measure Consumers' Judgments," *Harvard Business Review,* 53 (July-August), 107-117.

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science,* 12 (1), 1-27.

Guadagnoli, Edward and Wayne F. Velicer (1988), "Relation of Sample Size to the Stability of Component Patterns," *Psychological Bulletin,* 103, 265-275.

Haley, Russell and Ronald Gatty (1971), "The Trouble with Concept Testing," *Journal of Marketing Research,* 8 (2), 230-232.

Hoeffler, Steve (2003), "Measuring Preferences for Really New Products," *Journal of Marketing Research,* 40 (4), 406-420.

Holbrook, Morris B. and William L. Moore (1981), "Feature Interactions in Consumer Judgments of Verbal versus Pictorial Presentations," *Journal of Consumer Research,* 8 (1), 103-113.

Horn, J. L. (1965), "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, 30, 179-185.

Howe, T. (2004), "RCA Selecta Vision VideoDisc FAQ," available at www.faqs.org/faqs/rec-video/videodisc/selectavision.

Hughes, David G. and Jose L. Guerrero (1971), "Simultaneous Concept Testing with Computer-controlled Experiments," *Journal of Marketing,* 35 (January), 28-33.

Jarvis, W. Blair G. and Richard E. Petty (1996), "The Need to Evaluate," *Journal of Personality and Social Psychology,* 70 (1), 172-194.

Jones, L. (2005), *easyJet: The Story of Britain's Biggest Low Cost Airline*, London, Aurum Press.

Jöreskog, K. G. (1967), "Some Contributions to Maximum Likelihood Factor Analysis," *Psychometrika*, 32, 433-482.

Kaiser, H. F. (1960), "The Application of Electronic Computers to Factor Analysis," *Educational and Psychological Measurement*, 20, 141-151.

Kakkar, Pradeep and Richard J. Lutz (1975), "Toward a Taxonomy of Consumption Situations," *AMA Educator Conference Proceedings 1975,* 206-210.

Kalwani, Manohar U. and Alvin J. Silk (1982), "On the Reliability and Predictive Validity of Purchase Intention Measures," *Marketing Science,* 1 (3), 243-286.

Keeling, Kellie B. (2000), "A Regression Equation for Determining the Dimensionality of Data," *Multivariate Behavioral Research*, 35 (4), 457-468.

Klink, Richard R. and Gerard A. Athaide (2006), "An Illustration of Potential Sources of Concept-Test Error," *Journal of Product Innovation Management,* 23 (4), 359-370.

Krieger, Abba, Paul Green, Leonard Lodish, Jim D'Arcangelo, Chris Rothey and Paul Thirty (2003), "Consumer Evaluations of "Really New" Services: The TrafficPulse System," *Journal of Services Marketing*, 17 (1), 6-36.

Kristensson, Per, Anders Gustafsson and Trevor Archer (2004), "Harnessing the Creative Potential among Users," *Journal of Product Innovation Management,* 21 (4), 4-14.

Lee, Yikuan and Gina Colarelli O'Connor (2003), " The Impact of Communication Strategy on Launching New Products: the Moderating Role of Product Innovativeness," *Journal of Product Innovation Management*, 20 (1), 4-21.

Lees, Gavin and Malcolm Wright (2004), "The Effect of Concept Formulation on Concept Test Scores," *Journal of Product Innovation Management,* 21 (6), 389-400.

Leigh, James H. and Claude R. Martin, Jr. (1981), "A Review of Situational Influence Paradigms in Research," *AMA Review of Marketing,* 57-74.

Lewis, Ian M. (1984), "Do Concept Scores Measure the Message or the Method?" *Journal of Advertising Research,* 24 (1), 54-56.

Liberman, Nira and Yaacov Trope (1998), "The Role of Feasibility and Desirability Considerations in Near and Distant Future Decision: A Test of Temporal Construal Theory," *Journal of Personality and Social Psychology,* 75 (1), 5-18.

Loewenstein, George (1996), "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes,* 65 (March), 272-92.

Louviere, Jordan J., Herb Schroeder, Cathy H. Louviere and George G. Woodworth (1987), "Do the Parameters of Choice Models Depend on Differences in Stimulus Presentation: Visual versus Verbal Presentation?" *Advances in Consumer Research,* 14, 79-82.

Lynn, Michael and Judy Harris (1997), "The Desire for Unique Consumer Products: A New Individual Differences Scale," *Psychology & Marketing,* 14 (6), 601-616.

Mahajan, Vijay and Jerry Wind (1992), "New Product Models: Practice, Shortcomings and Desired Improvements," *Journal of Product Innovation Management,* 9 (2), 128-139.

Malkoc, Selin A. and Gal Zauberman (2006), "Deferring Versus Expediting Consumptions: The Effect of Outcome Concreteness on Sensitivity to Time Horizon," *Journal of Marketing Research,* Forthcoming.

Manning, Kenneth C., William O. Bearden and Thomas J. Madden (1995), "Consumer Innovativeness and the Adoption Process," *Journal of Consumer Psychology,* 4 (4), 329-346.

McCabe, George P. and David S. Moore (2005), *Introduction to the Practice of Statistics,* 5th edition, W. H. Freeman.

McCrae, Robert R. and Paul T. Costa, Jr. (1987), "Validation of the Five-Factor Model of Personality Across Instruments and Observers," *Journal of Personality and Social Psychology,* 52 (January), 81-90.

Midgley, David F. (1977), *Innovation and New Product Marketing.* London: Croom Helm.

Miller, James B., Norman T. Bruvold and Jerome B. Kernan (1987), "Does Competitive-set Information Affect the Results of Concept Tests?" *Journal of Advertising Research,* 27 (April/May), 16-24.

Mischel, Walter, Yuichi Shoda and Monica L. Rodriguez (1989), "Delay of Gratification in Children," *Science,* 244 (4907), 933-938.

Monga, Ashwani and Michael J. Houston (2006), "Fading Optimism in Products: Temporal Changes in Expectations about Performances," *Journal of Marketing Research,* Forthcoming.

Moore, William L. (1982), "Concept Testing," *Journal of Business Research,* 10, 279-294.

Moore, William L. and Morris B. Holbrook (1982), "On the predictive validity of joint-space models in consumer evaluations of new concepts," *Journal of Consumer Research,* 9, 206-210.

Moreau, Page C., Donald R. Lehmann and Arthur B. Markman (2001), "Entrenched Knowledge Structures and Consumer Response to New Products," *Journal of Marketing Research,* 38 (February), 14-29.

Morrison, Don G. (1979), "Purchase Intention and Purchase Behavior," *Journal of Marketing,* 43 (Spring), 65-74.

Nowland, Roger L. (1947), "Pre-Design Research as Applied to Product Development," in *American Society of Mechanical Engineers: Paper,* No. 47-A-57 (December).

Nunnally, Jum (1978), *Psychometric Theory,* 2nd Edition. New York: McGraw-Hill.

Ozanne, Julie L., Merrie Brucks and Dhruv Grewal (1992), "A Study of Information Search Behavior during the Categorization of New Products," *Journal of Consumer Research,* 18 (March), 452-463.

Ozer, Muammer (1999), "A Survey of New Product Evaluation Models," *Journal of Product Innovation Management,* 16, 77-94.

Page, Albert L. and Harold F. Rosenbaum (1992), "Developing an Effective Concept-Testing Program for Consumer Durables," *Journal of Product Innovation Management,* 9 (4), 267-277.

Paulus, Delroy L. (1991), "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes,* J.P. Robinson et al., eds. San Diego: Academic Press, 17-59.

Payne, John W., James R. Bettman and Eric J. Johnson (1992), "Behavioral Decision Research: A Constructive Processing Approach," *Annual Review of Psychology,* 43, 87-131.

Payne, John W., James R. Bettman and David A. Schkade (1999), "Measuring Constructed Preferences: Toward a Building Code," *Journal of Risk and Uncertainty,* 19 (1-3), 243-270.

Peter, Paul J. (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," *Journal of Marketing Research,* 16 (February), 6-17.

Rachlin, Howard (1995), "Self-Control: Beyond Commitment," *Behavioral and Brain Sciences,* 18 (1), 109-59.

Reidenbach, Eric R. and Sharon Grimes (1984), "How Concept Knowledge Affects Concept Evaluation," *Journal of Product Innovation Management,* 4, 255-266.

Rentz, Joseph O. (1987), "Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures," *Journal of Marketing Research,* 24 (February), 19-28.

---- (1988), "An Exploratory Study of the Generalizability of Selected Marketing Measures," *Journal of the Academy of Marketing Science,* 16 (1), 141-150.

Richins, Marsha L. and Scott Dawson (1992), "A consumer values orientation for materialism and its measurement: scale development and validation," *Journal of Consumer Research,* 19, 303-316.

Rook, Dennis W. and Robert J. Fisher (1995), "Normative Influences on Impulsive Buying Behavior," *Journal of Consumer Research,* 22, 305-313.

Ruiz-Primo, Maria A., Gail P. Baxter and Richard J. Shavelson (1993), "On the Stability of Performance Assessments," *Journal of Educational Measurement,* 30, 41-53.

Schindler, Robert M., Morris B. Holbrook and Eric A. Greenleaf (1989), "Using Connoisseurs to Predict Mass Tastes," *Marketing Letters,* 1 (December), 47-54.

Schoormans, Jan P.L., Roland J. Ortt and Cees J. P. M. de Bont (1995), "Enhancing Concept Test Validity by Using Expert Consumers," *Journal of Product Innovation Management,* 12, 153-162.

Shavelson, Richard J., Gail P. Baxter and Xiaohong Gao (1993), "Sampling Variability of Performance Assessments," *Journal of Educational Measurement,* 30, 215-232.

Slovic, Paul (1995), "The Construction of Preference," *American Psychologist,* 50 (5), 364-371.

Smead, Raymond J., James B. Wilcox and Robert E. Wilkes (1981), "How Valid Are Product Descriptions and Protocols in Choice Experiments?" *Journal of Consumer Research,* 8 (June), 37-41.

Snedecor, George W. and William G. Cochran (1989), *Statistical Methods*, Eighth Edition, Iowa State University Press.

Song, MX and MM Montoya-Weiss (1998), "Critical Development Activities for Really New versus Incremental Products," *Journal of Product Innovation Management*, 15(2), 124-135.

Srivastava, Raj K. (1980), "Usage-situational Influences on Perceptions of Product Markets: Response Homogeneity and its Implications for Consumer Research," *Advances in Consumer Research,* 5, 32-38.

Steenkamp, Jan-Benedict E. M. and Hans Baumgartner (1995), "Development and Cross-cultural Validation of a Short Form of CSI as a Measure of Optimum Stimulation Level," *International Journal of Research in Marketing,* 12 (2), 97-104.

Tauber, Edward M. (1972), "What Is Measured by Concept Testing?" *Journal of Advertising Research,* 12 (December), 35-37.

---- (1974), "How Market Research Discourages Major Innovation," *Business Horizons,* 17 (June), 22-26.

---- (1975), "Why Concept and Product Tests Fail to Predict New Product Results," *Journal of Marketing,* 39 (October), 69-71.

---- (1981), "Utilization of Concept Testing for New-product Forecasting: Traditional versus Multiattribute approaches," in *New-product Forecasting,* Y. Wind et al., eds. Lexington, MA: D. C. Heath: 169-178.

Taylor, James W., John J. Houlahan and Alan C. Gabriel (1975), "The Purchase Intention Question in New Product Development: A Field Test," *Journal of Marketing,* 39 (1), 90-92.

Thaler, Richard (1981), "Some Empirical Evidence on Dynamic Inconsistency," *Economics Letter,* 8 (3), 201-07.

Trebbi, George G. Jr. and Edward J. Flesch (1983), "Single Versus Multiple Concept Tests," *Journal of Advertising Research,* 23 (June/July), 21-26.

Trope, Yaacov and Nira Liberman (2000), "Temporal Construal and Time-Dependent Changes in Preference," *Journal of Personality and Social Psychology,* 79 (6), 876-89.

---- and ---- (2003), "Temporal Construal," *Psychological Review,* 110 (3), 403-421.

Urban, Glen L., Bruce D. Weinberg and John R. Hauser (1996), "Premarket Forecasting of Really New Products," *Journal of Marketing,* 60 (January), 47-60.

Urban, Glen L., John R. Hauser, William J. Qualls and Bruce D. Weinberg (1997), "Information Acceleration: Foundation and Lessons from the Field," *Journal of Marketing Research*, 34 (February), 143-153.

Viswanathan, Madhu (2005), *Measurement Error and Research Design*. Thousand Oaks, CA: Sage.

---- (1997), "Individual Differences in Need for Precision," *Personality and Social Psychology Bulletin,* 23 (7), 717-735.

Von Hippel, Eric (1986), "Lead Users: A Source of Novel Product Concepts," *Management Science,* 37 (2), 791-805.

Vriens, Marco, Gerard H. Loosschilder, Edward Rosbergen and Dick R. Wittink (1998), "Verbal versus Realistic Pictorial Representations in Conjoint Analysis with Design Attributes," *Journal of Product Innovation Management,* 15 (5), 455-467.

Warshaw, Paul R. (1980), "Predicting Purchase and Other Behaviors from General and Contextually Specific Intentions," *Journal of Marketing Research,* 17 (February), 26-33.

Webb, Noreen M., Jonah M. Schlackman and Brenda Sugrue (2000), "The Dependability and Interchangeability of Assessment Methods in Science," *Applied Measurement in Education,* 13, 277-301.

Wilton, Peter C. and Edgar A. Pessemier (1981), "Forecasting the Ultimate Acceptance of an Innovation: the Effects of Information," *Journal of Consumer Research,* 8 (September), 162-171.

Wind, Yoram (1973), "New Procedure for Concept Evaluation," *Journal of Marketing,* 37 (October), 2-11.

Wolpert, Henry W. (1980), "Why Conventional Automobile Styling Research May Become Obsolete," in *Advances in Consumer Research,* 7. Jerry C. Olson. Association for Consumer Research, Ann Arbor.

Zajonc, Robert (1968), "Attitudinal effects of mere exposure," *Journal of Personality and Social Psychology Monograph Supplement,* 9, 1-27.

Zaltman, Gerald and Melanie Wallendorf (1979), *Consumer Behavior: Basic Findings and Management Implications*. New York: Wiley.

Ziamou, Paschalina (Lilia) and Robert W. Veryzer (2005), "The Influence of Temporal Distance on Consumer Preferences for Technology-Based Innovations," *Journal of Product Innovation Management,* 22 (4), 336-346.

Zuckerman, Marvin (1979), *Sensation seeking: Beyond the optimal level of arousal.* Hillsdale, NJ: Lawrence Erlbaum.

Appendices

A-1 Materials for the Survey

1. Pre-notice Letter

April 25, 2005

Dear xxxxxx

A few days from now you will receive an email request to fill out a questionnaire for an important new product development research project being conducted by Professor Adam Finn and Ling Peng, a PhD student, from the University of Alberta School of Business. It concerns current new product development practice and the use of concept testing in your organization.

This survey is for managers who are responsible for new product development. If you are not the right person in your organization, we would appreciate it if you would forward the email to the right person. I am writing in advance because we have found many people like to know ahead of time that they will be contacted. The results of the study will have implications for the developers, suppliers and users of concept testing services and will provide us with important background information for our follow-up academic research we are planning on measurement issues in concept testing.

I would greatly appreciate it if you could take the time to complete the survey. In return, as a way of saying thanks for your help, we can promise you access to an aggregate level summary of the results of the study. It will only be with the generous help of people like you that our research can be successful.

Sincerely,

Ling Peng
University of Alberta School of Business
Edmonton, AB T6G 2R6
FAX 780-492-3325

2. Cover Letter

May 2, 2005
Dear xxxxxx

My name is Ling Peng. I am a Ph.D. student at the University of Alberta School of Business, doing my dissertation research on new product development and concept testing. I would like to invite you to participate in a research study about current concept testing practice that fulfills part of my dissertation requirements.

Answering the survey takes 15 to 20 minutes. Completing and returning the questionnaire will indicate consent has been given to use your responses for research purposes. There are no known risks to participating in the study. The identity of participants will be kept confidential. Answers used in the research will be anonymous. Results will only be released in a way that no individual's or firm's answers can be identified.

Answers will be used to better understand the practice, problems and desired improvements of concept testing. Further, only Professor Adam Finn, who is my dissertation supervisor, and I will have access to the raw data (without names). The survey responses will be printed out and stored in a locked file and destroyed once the data are coded and cleaned.

Other than the time burden, we anticipate no risks from participation in this research.

For your convenience there are two ways to answer this survey:

1. Open the attachment file, enter your responses and save, then email it back as an attachment to lpeng@ualberta.ca or adam.finn@ualberta.ca

2. Open the attachment file, print it out and return it with your answers to us via mail.

   Our mail address is:

   Ling Peng
   School of Business Building
   University of Alberta
   Edmonton, AB
   Canada T6G 2R6

If you require clarification of any of the questions used in the study or have any questions or concerns about this study and its findings, you may contact Ling Peng (myself) or Professor Finn. We would be happy to address any questions or concerns you may have.

Adam Finn	**Ling Peng**
University of Alberta	**University of Alberta**
adam.finn@ualberta.ca	**lpeng@ualberta.ca**
(780) 492-5369	**(780) 492-5435**

Any concerns about your treatment or rights as a research participant can be addressed to the Chair of the University of Alberta School of Business Research Ethics Board, who can be contacted at (780) 492-8443 or researchethicsboard@exchange.bus.ualberta.ca.

Attached is the Word file form of the survey. Your thoughts and experiences will be of great help to our research. Thank you very much!

Ling Peng
School of Business Building
University of Alberta
Edmonton, AB
Canada T6G 2R6

3. Thank You Letter

Dear xxxxxx May 9, 2005

About a week ago we sent you a survey via e-mail. We are asking product managers about new product development issues and the current concept testing practice in their organizations and the methods/models they used in their most recent concept testing project.

If you have already completed and returned the questionnaire, please accept our sincere thanks. If not, please do so today. We have contacted you and others now to obtain the many insights only product managers like you can provide.

We would greatly appreciate it if you could take a few moments to complete it. In return, and as a way of saying thanks for your help, we promise you access to an aggregate level summary of the study results. It is only with the generous help of people like you that our research can be successful.

In case the questionnaire and previous information about the survey has been deleted from your e-mail account, we have included them again.

Sincerely,

Ling Peng
School of Business Building
University of Alberta
Edmonton, AB Canada T6G 2R6

4. Survey Questionnaire

Please answer the following questions about your organization's new product development activities

S1. Please indicate the approximate number of new products your organization has introduced in each of the last 3 years (2002-2004). Of these, how many were radically new products?

	2002	**2003**	**2004**
The number of new products			
The number of radically new products			

S2. Please indicate the approximate number of concept tests conducted for your organization in each of the last 3 years (2002-2004). Of these, how many were for radically new products?

	2002	**2003**	**2004**
The number of concept tests			
The number of radically tests			

N1. Please indicate the proportion of new products for which each of the following new product development process activities were performed in the last three years. (Check one box in each row)

	Never	**In very few cases**	**In most cases**	**In all cases**
Formal new product idea generation process	☐	☐	☐	☐
Formal new product concept screening	☐	☐	☐	☐
Detailed market study prior to concept development testing	☐	☐	☐	☐
Detailed market study for market Identification, positioning and strategy	☐	☐	☐	☐
Business/financial analysis	☐	☐	☐	☐
Product development	☐	☐	☐	☐
Customer test of product prototype	☐	☐	☐	☐
Pre-market volume forecasting using prototype	☐	☐	☐	☐
Market test/trial sell	☐	☐	☐	☐
Market launch planning	☐	☐	☐	☐
Other	☐	☐	☐	☐

N2. And please indicate the importance of each activity. (Check one box in each row)

	Critical	**Important**	**Marginal**	**Not at all important**
Formal new product idea generation process	☐	☐	☐	☐
Formal new product concept screening	☐	☐	☐	☐
Detailed market study prior to concept development testing	☐	☐	☐	☐
Detailed market study for market Identification, positioning and strategy	☐	☐	☐	☐
Business/financial analysis	☐	☐	☐	☐
Product development	☐	☐	☐	☐
Customer test of product prototype	☐	☐	☐	☐
Pre-market volume forecasting using prototype	☐	☐	☐	☐
Market test/trial sell	☐	☐	☐	☐
Market launch planning	☐	☐	☐	☐
Other	☐	☐	☐	☐

N3. Please indicate which of the following models and/or methods your organization has used in its new product development in the last three years. (Check the boxes for all that apply)

☐ Ethnographic/observation usage research
☐ Focus group
☐ Limited rollout
☐ Concept tests
☐ Show tests and clinics
☐ Attitude and usage studies
☐ Traditional conjoint analysis
☐ Stated choice/preference models
☐ Delphi
☐ Quality function development (QFD)
☐ Home usage test
☐ Product life-cycle models
☐ Synectics
☐ Other (please specify):

N4. Please indicate your degree of satisfaction with the models and/or methods your organization has used in its new product development in the last three years. (Check a box in each row)

	Completely satisfied	Somewhat satisfied	Neither satisfied nor dissatisfied	Somewhat dissatisfied	Completely dissatisfied
Ethnographic/observation usage research	☐	☐	☐	☐	☐
Focus group	☐	☐	☐	☐	☐
Limited rollout	☐	☐	☐	☐	☐
Concept tests	☐	☐	☐	☐	☐
Show tests and clinics	☐	☐	☐	☐	☐
Attitude and usage studies	☐	☐	☐	☐	☐
Traditional conjoint analysis	☐	☐	☐	☐	☐
Stated choice/preference models	☐	☐	☐	☐	☐
Delphi	☐	☐	☐	☐	☐
Quality function development (QFD)	☐	☐	☐	☐	☐
Home usage test	☐	☐	☐	☐	☐
Product life-cycle models	☐	☐	☐	☐	☐
Synectics	☐	☐	☐	☐	☐
Other	☐	☐	☐	☐	☐

Now I would like to ask you some questions about your organization's ***most recent traditional or conjoint concept testing project*** *where the outcome has been determined and a resulting management decision has been made.*

C1. Please indicate the important objectives of this concept testing project. (Check the boxes for all that apply)

☐ Estimate the concept's market potential
☐ Eliminate the poor concept(s)
☐ Generate an estimate of the sales or trial rate
☐ Develop further the original idea
☐ Help identify the highest potential customer segment
☐ Provide diagnostic information
☐ Identify the value of concept features
☐ Other (please specify):

C2. Please indicate what type of product or service you were concept testing (Check a box).

- ☐ Industrial
- ☐ Service
- ☐ Package Goods
- ☐ Durables
- ☐ Others

C3. Please indicate the newness of the product concept (Check one box).

- ☐ Radically new to the market
- ☐ Incrementally new to the market
- ☐ Improvement to an existing product
- ☐ Repositioning of an existing product
- ☐ Extension of an existing product
- ☐ Other (please specify):

C4. Please indicate the number of concepts that were evaluated in that testing project.

Concepts

C5. Was it a monadic test or a comparative test?

- ☐ Monadic test
- ☐ Comparative test

C6. Did the concept description include pricing information?

- ☐ Pricing information included
- ☐ No pricing information

C7. Concepts can be presented in varied ways. Please indicate the form of concept presentation used in your most recent concept testing project. (Check one box)

- ☐ Stripped description (simple, factual, non-emotional written description of the idea)
- ☐ Embellished description (advertising format, promotional, full description)
- ☐ Stripped with visual representation
- ☐ Rough mock advertisement presentation
- ☐ Fully finished advertisement presentation
- ☐ Other (please specify):

C8. Which of the following types of questions did you use to measure the respondents' reaction to the concept in your most recent concept test project? (Check the boxes for all that apply)

Unstructured/qualitative open-ended evaluations of:

☐ Problem solving ability
☐ Liking
☐ Disliking
☐ Believability
☐ Uniqueness
☐ Comparison with current offering
☐ Comparison with expectation
☐ Purchase intent
☐ Others (please specify):

Structured/quantitative rating scale evaluations of:

☐ Problem solving ability
☐ Liking
☐ Disliking
☐ Believability
☐ Uniqueness
☐ Comparison with current offering
☐ Comparison with expectation
☐ Purchase intent
☐ Others (please specify):

C9. For each of the structured/quantitative questions, what number of scale points did your response scale have?

Questions used	**Number of scale points**
Problem solving ability	
Liking	
Believability	
Uniqueness	
Comparison with current offering	
Comparison with expectation	
Purchase intent	
Other question:	
Other question:	

C10. What type of response scale format did you use for rating scale questions? (Check one box)

☐ Numerical categories only
☐ Categories with verbal end-point labels only
☐ Categories with verbal end-point and mid-point labels only
☐ Categories with all point labeled
☐ Both numerical categories and verbal labels
☐ Continuous response scale (e.g., line, thermometer)
☐ Other (please specify):

C11. Which of the following best describes how you assessed the outcome of the concept test for the quantitative questions you used? (Check the boxes for all that apply)

☐ Percent top-box score
☐ Percent top-2-box scores
☐ Rating scale mean
☐ Rating scale median
☐ Comparison with a company specific norm
☐ Comparison with an industry specific norm
☐ Comparison with a research supplier norm
☐ Other (please specify):

C12. What criteria did you use to select the respondents used in the concept test? (Check the boxes of all that apply)

☐ Product class usage
☐ Specific product usage
☐ Innovativeness
☐ Lead user criteria
☐ Influentials/market maven criteria
☐ Demographics
☐ Lifestyle group membership
☐ Other (please specify):

C13. How many respondents evaluated each concept in the concept-testing project?

Respondents

C14. Which of the following best describes the data collection method used for the concept-testing project? (Check one box)

- ☐ Focus groups
- ☐ In-home interviews
- ☐ Mall intercepts/Central location
- ☐ Telephone interviews
- ☐ Mail survey
- ☐ On-line survey
- ☐ Other (please specify):

C15. What percentage of the concept testing projects you have performed in the last three years (2002-2004) were performed on-line?

Percent on-line

C15a. What are the primary reasons for you to choose on-line concept testing?

C16. How satisfied are you with the predictive performance of your organization's **current approach to concept testing**. (By predictive performance I mean reliability and validity - the degree to which concepts that score well at this stage also score well in subsequent concept tests or perform well in the market).

- ☐ Completely satisfied
- ☐ Somewhat satisfied
- ☐ Neither satisfied nor dissatisfied
- ☐ Somewhat dissatisfied
- ☐ Completely dissatisfied

C17. What are the primary problems you see with concept testing?

C18. Under what circumstances do concept tests work best?

C19. Under what circumstances are concept tests least likely to work well?

C20. What, if any, changes do you expect to see in the way concept tests are conducted in the next five years?

Please answer the following questions to be used for classification purposes

F1. Please indicate the industry type of your company.

☐ Service
☐ Package goods
☐ Durables
☐ Industrial
☐ Other

F2. Please indicate the percent of your organization's current sales that are attributable to products that are less than 3 years old?

Percent

F3. Please indicate whether the market focus of your company is domestic or global?

☐ Domestic
☐ Global

F4. Please indicate whether the product development focus of your company is domestic or global?

☐ Domestic
☐ Global

Our survey ends here, thank you for your time and cooperation! If there is anything else you would like to tell us about the role of concept testing please do so here.

☐ Check this box and provide a contact address to request a summary of survey results

Your contact address:

5. Letter to Marketing Research Firms for Requesting Information

Dear Sir or Madam,

I am Ling Peng, a PhD student in Marketing, University of Alberta School of Business.

I am doing my dissertation research on new product development and concept testing. I am approaching some research consultants like you for more information about current new product development practice and the use of concept testing in your firms. The results of the study will have implications for the developers, suppliers and users of concept testing services and will provide us with important background information for our follow-up academic research we are planning on measurement issues in concept testing.

Would you please tell me (1) does your company offer product/concept testing services? (2) What types of concept testing do you provide (quantitative or qualitative)? (3) Do you have track record of reliability and validity for the concepts tested earlier?

Thanks for your information. It will only be with the generous help of people like you that our research can be successful.

Ling Peng
lpeng@ualberta.ca
238 RH Michener Park
Edmonton, Alberta, Canada
T6H 4M5

A-2 Materials for the Pretest

1. Pretest Recruitment

Subject:
[Institute for Online Consumer Studies] Invitation to participate in a web-based study

Dear member of the Institute for Online Consumer Studies (IOCS) participant panel,

We invite you to participate in a web-based study about new consumer appliances. This research is being conducted by Ling Peng and Adam Finn at the University of Alberta.

Completing this study takes about 30 minutes. You will be asked to rate 20 descriptions of consumer appliances with respect to product newness and clearness of the descriptions.

To be eligible for this study, you must
- have some interest in consumer appliances, and
- own and/or use some particular consumer appliances, and
- have a valid e-mail address linked to an active paypal account.

You will be paid either USD 5.00 (US $) or CAD $6.00 (Canadian $) - you can choose the currency - for your participation in this study. Only those who complete the study and provide a valid email address linked to an active Paypal account will be paid. The payment will take place within 7-10 days after the completion of the study.

In order to participate in the study, please click on the following link or copy and paste it in your web browser's address bar:

http://www.iocs.org/npt33.cfm.

A limited number of participants is needed for this study. The study will be available until it has been completed by the required number of participants.

Sincerely,

IOCS team
http://www.iocs.org

To unsubscribe from the IOCS panel, go to http://www.iocs.org/panel.cfm.

2. Product Descriptions

1. Dental Water Jet - This Dental Water Jet is designed to flush food debris from hard-to-reach places between teeth or around dental implants, orthodontic braces, and crowns. The easy-fill, translucent tank holds water, dental rinse, or mouthwash. This Water Jet Model has 7-speed, variable-flow water pressure and comes with four jet tips.

2. Flosser - This flosser makes flossing easier and less messy. With a simple push of a button, this flosser gently vibrates to reach between your teeth and below your gum line to remove plaque in areas your toothbrush can't. Features include: Choice of two interchangeable attachments - power ergonomic flosser and soft mint dental pick; Single-use refill attachments; Easy and convenient size lets you floss with one hand; Handy 4-inch tall compact design fits in the palm of your hand.

3. Ear Thermometer - This ear thermometer is designed to gently take a temperature in just seconds by measuring the heat generated by the eardrum and surrounding tissue. Features include: Temperature range: 34 - 42.2C; Temperature range: 93.2 - 180.0F; Ready Beep; Small, soft tip; Auto-off after 120 sec; Easy-to-read LCD display; Lens filter detector; Lens filter ejector; Protective cap.

4. Smartphone - This smartphone is designed to combine phone, computer, and entertainment functions into a single device. The unit includes an MP3 player designed to provide CD-quality stereo sound. The media player shows images in jpg, bmp, tif, and png formats and shows video in mov, avi, aud, and wmf formats.

5. Smartphone - This smartphone combines the functions of a phone and a personal digital assistant (PDA) with mobile office and e-mail capabilities. It is also a digital and video VGA camera, as well as an MP3 music and MPEG4 video player. The device allows the user to input information and control the phone using a full qwerty keyboard inside the flip, or, with the flip closed, by use of a jog dial on the side of the phone. The unit offers handwriting recognition and predictive text input capabilities.

6. Portable Heater - Because Mother Nature isn't always cooperative, here's a portable heater on wheels that makes it easier to put the light and heat where you need it. The DCS light and heater come in your choice of stainless steel and five textured-color finishes. The propane gas heater has a 40,000-BTU burner system with 100-percent safety shutoff, and heats an area up to 20 feet.

7. Steam Iron - This steam iron features 1,200 watts of power and a continuous steam system that adapts the steam level to different fabrics. With 17 steam vents for consistent and thorough steam distribution, the iron provides a burst of steam when you need it. A stainless-steel soleplate is helpful for smooth gliding over all fabrics. A large transparent water tank allows for longer use without refilling. A swivel joint keeps the cord up and out of the way, making in comfortable for left- and right-hand users.

8. Stairmaster - This product is designed to deliver exceptional results by achieving the same cardiovascular workout you get on a treadmill, plus a superior lower-body muscular workout - working your heart, lungs and muscles without jarring your joints. Features include: Revolving staircase speed may be varied from 24-162 steps per minute, Easy-to-read LCD console; Motivating programs include a nationally recognized fitness test to gauge individual progress and a custom firefighter test (C.P.A.T.); Includes Polar® compatible telemetry Heart Rate monitoring.

9. GPS - This GPS has advanced points of interest database. Easy to use touch-screen interface. Comes with sunction cup windshield mount DC vehicle power adapter, AC power supply, and data cable. Turn-by-turn guidance with voice prompting. Comprehensive mapping of U.S. and Canada on integrated a 20 GB hard drive. Exceptional GPS accuracy and reliability.

10. Hair Cutting Tool - If you can comb hair, you can cut hair! This product is said to be the world's first mistake-free hair cutting tool. This technology is designed to made to work with all types & textures of hair; this product helps extend the look of a professional haircut by weeks or even months. Highlights include: Cuts just a few hairs at a time for a smooth, even style; Cuts without streaks, gashes or cut lines; Effortlessly blends, tapers and shapes; Uses smooth, even strokes in a downward motion, in the direction that the hair grows.

11. Oral Care System - This product is said to be the most advanced oral care system inspired by dental professionals for ultimate cleaning, superior whitening & polishing for healthy teeth and gums. It is the only toothbrush system with unique 3D Excel technology - 40,000 in-and-out pulsations per minute gently loosen plaque, and 8,800 side-to-side oscillations sweep plaque away to help keep your mouth at its healthiest.

12. Pedometer - This pedometer is engineered with sensing technology utilizing two acceleration sensors. The sensors and unique data analysis algorithm allow the unit to be placed or worn in multiple positions. The Model HJ-112 is designed to measure steps while worn in a jacket or pants pocket, or even while in a bag. The unit calculates calories burned by factoring the walker's weight, stride distance, and number of steps. It displays calories burned as well as distance walked. Data can be stored and reviewed by users.

13. Personal Computer - This product is said to be the smallest, high-performance Windows XP computer with complete PC functionality. The unit's main feature is its modular design. When the screen is open, users can input data via a thumb keyboard with mouse buttons and a thumbwheel. When the screen is closed, the unit turns into a tablet-style PDA, complete with digital pen for input. If connected to the unit's docking cable, the uPC's capabilities are expanded to include video, Ethernet, and audio line out, and it can be connected to LAN or high-speed data networks.

14. Portable Media Center - This portable media center is a Windows Mobile-based portable device that makes it easy to store and play up to 80 hours of TV and movies, 5,000 songs and your most prized photos. Features MP3, WMA, WMV and JPEG playback; 3.5" TFT LCD screen; 20GB hard disk drive; USB 2.0 interface; and much more. System Requirements: Windows CP and up. Includes A/V-out cable, earphones, USB 2.0 cable and charger, install disk, carrying case and user manual.

15. MP3 Player - This 256MB MP3 Player is ultra light and portable and still packed with all the expected features offered on most players. Enjoy 10 hours of playback time from 1 AAA battery. Features: USB interface; EQ with 4 presets; MWA and MP3 compatible; 256 MB of internal memory; backlit LCD display with ID3 tag support; PC and Mac compatibility; 5 repeat modes; USB cable; installation CD; user's guide; carrying clip; and stereo earphones.

16. Digital Camera - This Super-Slim Digital Camera with Rotating Lens is a video camera, still camera, and it plays music--all in one! Features a 1.5" TFD LCD display; 2.11 total megapixels CCD; 2-1/2,000 second shutter speed; F4 aperture; 1x-4x digital still zoom; normal/macro focus; 1600x1200 pixels image size; black/white and Sepia color; multi-mode flash; three frames per second consecutive shooting; SD memory card; QuickTime Motion JPEG video compression (320x240 pixels); and WMA/MP3/AAC music.

17. Stereo Radio - Enjoy radio programming to go with this ultra-slim and compact Personal AM/FM stereo radio. Features up to 20 memory presets, DBBS-Dynamic bass boost, lightweight stereo earphones and a built-in belt clip. Digital AM/FM stereo radio also features an LED power on/off indicator. 1-3/4Wx3/4Hx4-1/4D".

18. Clock Radio - Enjoy an extra sense of security with this emergency alert and clock radio combo. The alert system receives all NOAA weather bands to warn you about emergency conditions and severe weather by automatically triggering a loud siren and flashing warning light. The unit doubles as an AM/FM radio with sleep timer and everyday alarm clock with snooze function. Rest easy with the standby light that indicates the radio is functioning properly. Large push-button.

19. Bicycle - Designed for leisure, commuting and touring; full size bike feel and drivability with the practicality of a stretching bicycle. Larger frame is well suited for taller and heavier riders; can be stretched or compressed in seconds with no tools required. Standard equipment includes 6-speed gearing; Shimano derailleur; kickstand, fenders, 2-tone comfort saddle, front and rear V-brakes bell, full set of reflectors and folding pedals.

20. Go-kart - This Ground Force electric powered go-kart offers whisper-quiet operation at speeds up to 12 mph. The solid steel construction offers rugged durability and safety. Other features include hand-controlled variable speed accelerator, molded aluminum wheels with solid rubber tires, high-torque motor and should seat belt. Runs up to 45 minutes on a full charge; charges in 4-6 hours.

3. Newness and Clearness Measures

Newness Measures

The newness scale was adapted from Lee and O'Connor (2003)

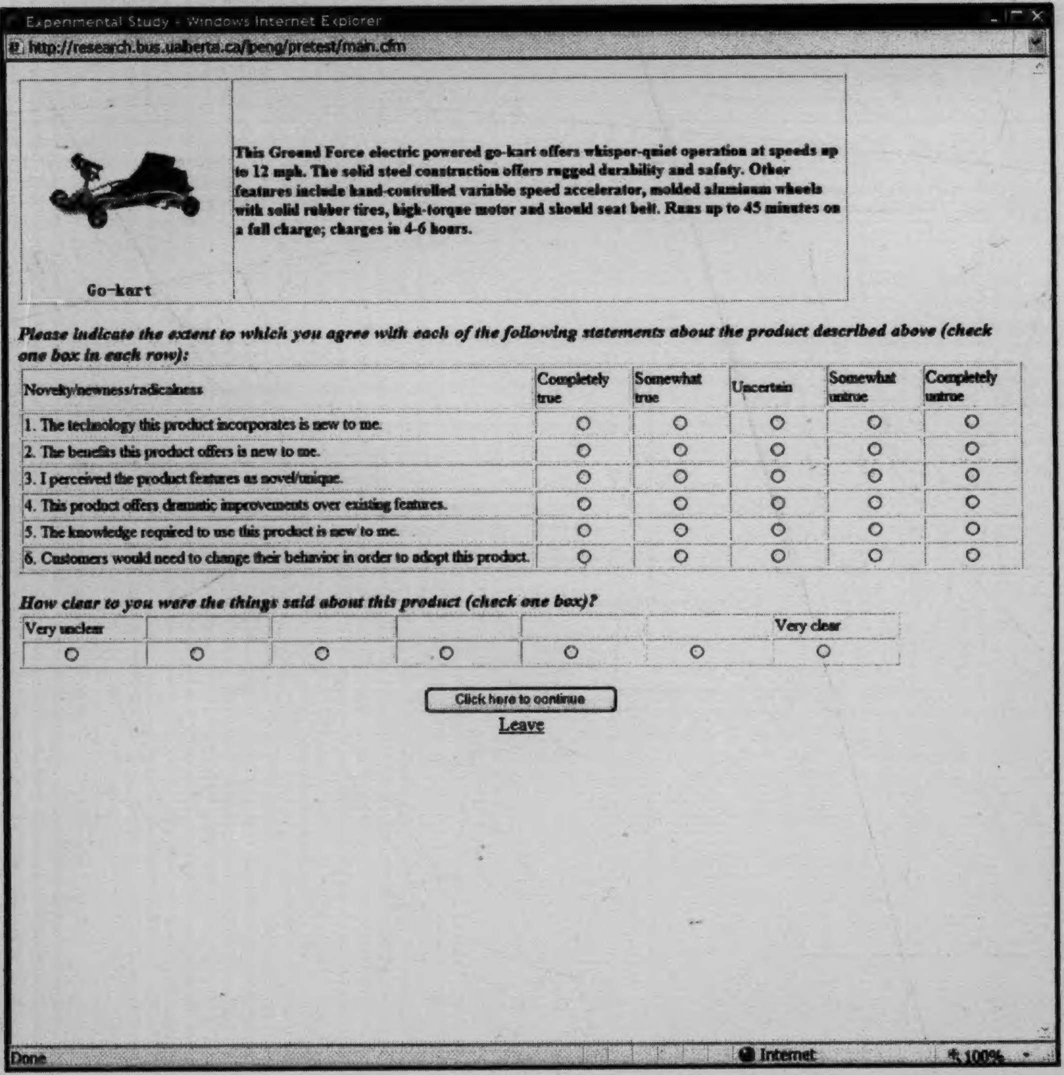
Please indicate the extent to which you agree with each of the following statements about the product described above (Completely true = 5; Somewhat true = 4; Uncertain = 3; Somewhat untrue = 2; Completely untrue = 1)

1. The technology this product incorporates is new to me.
2. The benefit this product offers is new to me.
3. I perceived the product features as novel/unique.
4. This product offers dramatic improvements over existing features.
5. The knowledge required to use this product is new to me.
6. Customers would need to change their behavior in order to adopt this product.

Clearness Measure

How clear to you were the things said about this product? (Very unclear = 1; Very clear = 7)

4. An Example of Response Screen Used in Pretest



Experimental Study - Windows Internet Explorer
http://research.bus.ualberta.ca/peng/pretest/main.cfm
Go-kart
Done
Internet
100%

This Ground Force electric powered go-kart offers whisper-quiet operation at speeds up to 12 mph. The solid steel construction offers rugged durability and safety. Other features include hand-controlled variable speed accelerator, molded aluminum wheels with solid rubber tires, high-torque motor and should seat belt. Runs up to 45 minutes on a full charge; charges in 4-6 hours.

Please indicate the extent to which you agree with each of the following statements about the product described above (check one box in each row):

Novelty/newness/radicalness	Completely true	Somewhat true	Uncertain	Somewhat untrue	Completely untrue
1. The technology this product incorporates is new to me.	○	○	○	○	○
2. The benefits this product offers is new to me.	○	○	○	○	○
3. I perceived the product features as novel/unique.	○	○	○	○	○
4. This product offers dramatic improvements over existing features.	○	○	○	○	○
5. The knowledge required to use this product is new to me.	○	○	○	○	○
6. Customers would need to change their behavior in order to adopt this product.	○	○	○	○	○

How clear to you were the things said about this product (check one box)?

Very unclear						Very clear
○	○	○	○	○	○	○

Click here to continue

Leave

A-3 Pretest Results

Table A-1
Newness and clearness Scores

Concept	Newness Score (Mean of six items)	Clearness	Chosen Appliances
Ear Thermometer	2.1589	6.3953	Minor
Stereo Radio	2.5930	6.3721	Minor
Dental Water Jet	2.6977	6.2971	Minor
MP3 Player	2.7364	6.0000	Minor
Oral Care System	2.7636	6.1163	
Steam Iron	2.8450	6.2791	Minor
Clock Radio	2.8450	6.3023	
GPS	3.1124	5.7674	
Pedometer	3.1318	5.9302	
Go-kart	3.1473	6.0645	
Portable Heater	3.1783	6.0233	
Bicycle	3.1938	6.0930	
Stairmaster	3.2171	6.1395	
Flosser	3.4186	6.1860	
Smartphone 1	3.4574	5.7442	
Hair Cutting Tool	3.4884	6.0930	Major
Digital Camera	3.6008	5.7907	Major
Smartphone 2	3.6473	6.0233	Major
Portable Media Center	3.8682	5.8140	Major
Personal Computer	3.9302	5.8140	Major

A-4 Materials for the Main Study

1. The First Wave Study Recruitment

Subject:
[Institute for Online Consumer Studies] Invitation to participate in a web-based study

Dear member of the Institute for Online Consumer Studies (IOCS) participant panel,

We invite you to participate in a web-based study (NPT34) about new consumer appliances. This research is being conducted by Ling Peng and Adam Finn at the University of Alberta.

The study consists of three waves. This is the first wave. Besides this wave, you will be recontacted to complete the following two waves at the end of October and November. Completing this wave takes about 30 minutes. You will be asked to evaluate 10 descriptions of consumer appliances and answer a few of opinion questions.

To be eligible for this study, you must
- have some interest in consumer appliances, and
- own and/or use some particular consumer appliances, and
- not ever participate in the NPT33 study conducted on September 20, 2005, and
- have a valid e-mail address linked to an active paypal account.

You will be paid either USD 5.00 (US $) or CAD $6.00 (Canadian $) - you can choose the currency - for your participation in the first wave and will be paid either USD 7.00 or CAD 8.00 each time you participate in the second and the third wave. Further, you will be offered either USD 5.00 or CAD 6.00 bonus at the end of the study if you complete all three waves. That means you will get either USD 24.00 or CAD 28.00 in total for the whole NPT34 study. Please note that your responses will be examined to see whether the task was been completed conscientiously to receive the bonus because we don't want to pay out to someone who responds by automatically entering the same or random answers. Only those who complete the survey questions and provide a valid email address linked to an active Paypal account will be paid. The payment will take place within 7-10 days after the completion of each wave.

In order to participate in the study, please click on the following link or copy and paste it in your web browser's address bar: http://www.iocs.org/npt34.cfm.

A limited number of participants is needed for this study. The study will be available until it has been completed by the required number of participants.

Sincerely,
IOCS team
http://www.iocs.org
To unsubscribe from the IOCS panel, go to http://www.iocs.org/panel.cfm.

2. The Second Wave Study Recruitment

Subject:
[Institute for Online Consumer Studies] Invitation to participate in the 2nd wave of the NPT34 study

Dear participant of the NPT34 study,

We invite you to participate in the second wave of a web-based study (NPT34) about new consumer appliances. This research is being conducted by Ling Peng and Adam Finn at the University of Alberta.

About a month ago you participated in the first wave of the NPT34 study. We were asking for your opinion about some consumer appliances concepts. Please accept our sincere thanks.

Here is the second wave of the NPT34 study. Completing this wave takes about 30 minutes. You will be asked to evaluate 10 descriptions of some consumer appliances (similar to the ones in the first wave) and answer a few of other opinion questions. Please note that you will be recontacted to complete the third wave at the end of November.

To be eligible for this study, you must have participated in the first wave of the NPT34 study and have a valid e-mail address linked to an active paypal account.

You will be paid either USD 7.00 or CAD 8.00 - you can choose the currency - each time you participate in the second and the third wave. Further, you will be offered either USD 5.00 or CAD 6.00 bonus at the end of the study if you complete all three waves. That means you will get either USD 24.00 or CAD 28.00 in total for the whole NPT34 study. Please note that your responses will be examined and the completion time will be recorded automatically to see whether the task was been completed conscientiously to receive the bonus because we don't want to pay out to someone who responds by automatically entering the same or random answers. Only those who complete the survey questions and provide a valid email address linked to an active Paypal account will be paid. The payment will take place within 7-10 days after the completion of each wave.

In order to participate in the second wave of the NPT34 study, please click on the following link or copy and paste it in your web browser's address bar:
http://www.iocs.org/npt34.cfm.
Please complete the study as soon as possible. The study will be available until 3PM November 3.

Sincerely,

IOCS team
http://www.iocs.org
To unsubscribe from the IOCS panel, go to http://www.iocs.org/panel.cfm.

3. The Third Wave Study Recruitment

Subject:
[Institute for Online Consumer Studies] Invitation to participate in the 3rd wave of the NPT34 study

Dear participant of the NPT34 study,

We invite you to participate in the third wave of a web-based study (NPT34) about new consumer appliances. This research is being conducted by Ling Peng and Adam Finn at the University of Alberta.

Please accept our sincere thanks for your participation in the previous waves of the study. This is the last wave of the NPT34 study. Completing this wave takes about 30 minutes. Similar to the first two waves, you will be asked to evaluate 10 descriptions of some consumer appliances and answer a few opinion questions.

To be eligible for this study, you must have participated in the first wave and/or second wave of the NPT34 study and have a valid e-mail address linked to an active paypal account.

You will be paid either USD 7.00 or CAD 8.00 - you can choose the currency – for your participation in the third wave. Further, you will be offered either USD 5.00 or CAD 6.00 bonus if you complete all three waves at the end of the study. Please note that your responses will be examined and the completion time will be recorded automatically to see whether the task was been completed conscientiously to receive the bonus because we don't want to pay out to someone who responds by automatically entering the same or random answers. Only those who complete the survey questions and provide a valid email address linked to an active Paypal account will be paid. The payment will take place within 7-10 days after the completion of each wave.

In order to participate in the third wave of the NPT34 study, please click on the following link or copy and paste it in your web browser's address bar:

http://www.iocs.org/npt34.cfm (Murat, please doublecheck this link for sure)

Please complete the study as soon as possible. The study will be available for three days.

Sincerely,

IOCS team
http://www.iocs.org
To unsubscribe from the IOCS panel, go to http://www.iocs.org/panel.cfm.

4. An Example of Reminder Letter

Dear Participants of the NPT34 study,

Please accept our sincere thanks for your participation in the first wave of the NPT34 study. About three days ago we sent you the link to the second wave of the NPT34 study via e-mail. In this wave, you will be asked to evaluate 10 descriptions of some consumer appliances (similar to the ones in the first wave) and answer a few of other opinion questions. Please note that the whole NPT34 study consists of three waves so you will be recontacted to complete the third wave at the end of this November.

You will be paid either USD 7.00 or CAD 8.00 - you can choose the currency - each time you participate in the second and the third wave. Further, you will be offered either USD 5.00 or CAD 6.00 bonus at the end of the study if you complete all three waves.

If you have already completed the second wave, please accept our thanks. If not, please do so as soon as possible. We will leave the link open for two more days. In order to participate in the second wave of the NPT34 study, please click on the following link or copy and paste it in your web browser's address bar:

http://www.iocs.org/npt34.cfm.

Sincerely,

Researchers of the NPT34 study
Ling Peng

5. Concept Evaluation Items

1. Liking
   Now I would like you to think about how much you would like to have this product?
   (Definitely not like = 1; Definitely like = 7)

2. Importance
   How important do you think the functions and features of this product are?
   (Not at all important =1; Very important=7)

3. Uniqueness
   How would you rate this product in terms of being unique from the products currently sold?
   (Not at all unique = 1; Very unique = 7)

4. Problem Solving
   How sure are you that this product would solve a problem for you?
   (Absolutely sure it will not = 1; Absolutely sure it will = 7)

5. Believability
   How believable to you were the things said about this product?
   (Don't believe it = 1; Firmly believe it = 7)

6. Purchase intention
   Assuming this product was available in a store where you shop, how likely would you be to buy it?
   (Definitely not buy = 1; Definitely would buy = 7)

6. Scales used for Personality Traits

Note: ® - Negative wording item

Need for Precision (Viswanathan 1997)

1. I enjoy tasks that require me to be exact
2. Vague descriptions leave me with the need for more information
3. I have a rough rather than exact idea of my opinions on various issues ®
4. I do not find it interesting to learn precise information ®
5. Thinking is enjoyable when it does not involve exact information ®
6. I tend to put things into broad categories as much as possible ®
7. I don't see the point in trying to discriminate between slightly different alternatives ®
8. I like to express myself precisely even when it is not necessary
9. I think approximate information is acceptable whereas exact information is not necessary ®
10. I am satisfied with information as long as it is more or less close to the facts ®
11. I am satisfied with my knowledge about issues as long as I am in the ballpark ®
12. I like tasks which require me to look for small differences between things
13. I like to use the precise information that is available to make decisions

Need to Evaluate Scale (Jarvis and Petty 1996)

1. I form opinions about everything
2. I prefer to avoid taking extreme positions ®
3. It is very important to me to hold strong opinions
4. I want to know exactly what is good and bad about everything
5. I often prefer to remain neutral about complex issues ®
6. If something does not affect me, I do not usually determine if it is good or bad ®
7. I enjoy strongly liking and disliking new things
8. There are many things for which I do not have a preference ®
9. It bothers me to remain neutral
10. I like to have strong opinions even when I am not personally involved
11. I have many more opinions than the average person
12. I would rather have a strong opinion than no opinion at all
13. I pay a lot of attention to whether things are good or bad
14. I only form strong opinions when I have to ®
15. I like to decide that new things are really good or really bad
16. I am pretty much indifferent to many important issues ®

Buying Impulsiveness (Impulsivity) (Rook and Fisher 1995)

1. I often buy things spontaneously
2. "Just do it" describes the way I buy things
3. I often buy things without thinking
4. "I see it, I buy it" describes me
5. "Buy now, think about it later" describes me
6. Sometimes I feel like buying things on the spur-of-the-moment
7. I buy things according to how I feel at the moment
8. I carefully plan most of my purchases ®
9. Sometimes I am a bit reckless about what I buy

Domain- or product category-specific innovativeness (Goldsmith and Hofacker 1991)

1. In general, I am among the first in my circle of friends to buy a new electronics when it appears.
2. If I heard that a new electronics was available in the store, I would not be interested enough to buy it. ®
3. Compared to my friends I own a lot of electronics.
4. In general, I am the first in my circle of friends to know the titles/brands of the latest electronics.
5. I will not buy a new electronics if I haven't heard/tried it yet. ®
6. I do not like to buy electronics before other people do. ®

Consumer innovativeness (Manning, Bearden, and Madden 1995)

- CIJM Items

1. Prior to purchasing a new brand, I prefer to consult a friend that has experience with the new brand. ®
2. When it comes to deciding whether to purchase a new service, I do not rely on experienced friends or family members for advice.
3. I seldom ask a friend about his or her experiences with a new product before I buy the new product.
4. I decide to buy new products and services without relying on the opinions of friends who have already tried them.
5. When I am interested in purchasing a new service, I do not rely on my friends or close acquaintances that have already used the new service to give me information as to whether I should try it.
6. I do not rely on experienced friends for information about new products prior to making up my mind about whether or not to purchase.

- CNS Items

1. I often seek out information about new products and brands.
2. I like to go to places where I will be exposed to information about new products and brands.
3. I like magazines that introduce new brands.

4. I frequently look for new products and services.
5. I seek out situations in which I will be exposed to new and different sources of product information. ®
6. I am continually seeking new product experiences.
7. When I go shopping, I find myself spending very little time checking out new products and brands.
8. I take advantage of the first available opportunity to find out about new and different products.

Uniqueness: Desire for unique consumer products (Lynn and Harris 1997)

1. I am very attracted to rare objects.
2. I tend to be a fashion leader rather than a fashion follower.
3. I am more likely to buy a product if it is scarce.
4. I would prefer to have things custom-made than to have them ready-made.
5. I enjoy having things that others do not.
6. I rarely pass up the opportunity to order custom features on the products I buy.
7. I like to try new products and services before others do.
8. I enjoy shopping at stores that carry merchandise that is different and unusual.

Involvement (Study Task) (Ozanne, Brucks and Grewal 1992)

1. I wanted to do a good job in this study.
2. I did not care about performance of this study. ®
3. The study was enjoyable.
4. The study was interesting.
5. I do not recommend participation in this study. ®

Exploratory Buying Behavior Tendencies (Baumgartner and Steenkamp 1996)

- EAP items

1. Even though certain food products are available in a number of different flavors, I tend to buy the same flavors. ®
2. I would rather stick with a brand I usually buy than try something I am not very sure of. ®
3. I think of myself as a brand-loyal consumer. ®
4. When I see a new brand on the shelf, I am not afraid of giving it a try.
5. When I go to a restaurant, I feel it is safer to order dishes I am familiar with. ®
6. If I like a brand, I rarely switch from it just to try something different. ®
7. I am very cautious in trying new or different products. ®
8. I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchases.
9. I rarely buy brands about which I am uncertain how well they perform. ®
10. I usually eat the same kind of foods on a regular basis. ®

- EIS items

1. Reading mail advertising to find out what's new is a waste of time. ®

2. I like to go window-shopping and find out about the latest styles.
3. I get very bored listening to others about their purchases. ®
4. I generally read even my junk mail just to know what it is about.
5. I don't like to shop around just out of curiosity. ®
6. I like to browse through mail order catalogs even when I don't plan to buy anything.
7. I usually throw away mail advertisements without reading them. ®
8. I like to shop around and look at displays.
9. I don't like to talk to my friends about my purchases. ®
10. I often read advertisements just out of curiosity.

Change Seeking Index: CSI Short Form (Steenkamp and Baumgartner 1994)

1. I like to continue doing the same old things rather than trying new and different things. ®
2. I like to experience novelty and change in my daily routine.
3. I like a job that offers change, variety, and travel, even if it involves some danger.
4. I am continually seeking new ideas and experiences.
5. I like continually changing activities.
6. When things get boring, I like to find some new and unfamiliar experience.
7. I prefer a routine way of life to an unpredictable one full of change. ®

Preference for Consistency: PFC (Cialdini, Trost, and Newsom 1995)

1. It is important to me that those who know me can predict what I will do.
2. I want to be described by others as a stable, predictable person.
3. The appearance of consistency is an important part of the image I present to the world.
4. An important requirement for any friend of mine is personal consistency.
5. I typically prefer to do things the same way.
6. I want my close friends to be predictable.
7. It is important to me that others view me as a stable person.
8. I make an effort to appear consistent to others.
9. It doesn't bother me much if my actions are inconsistent. ®

Social desirability (Richins and Dawson 1992)

1. I sometimes feel resentful when I don't get my way. ®
2. I am always careful about my manner of dress.
3. My table manners at home are as good as when I eat out in a restaurant.
4. There have been times when I felt like rebelling against people in authority even though I knew they were right. ®
5. I am always willing to admit it when I have made a mistake.
6. I sometimes try to get even rather than forgive and forget. ®
7. I am always courteous, even to people who are disagreeable.
8. I have never been irked when people expressed ideas very different from my own.
9. I am sometimes irritated by people who ask favors of me. ®
10. I have never deliberately said something that hurt someone's feelings.

7. Validation Tasks Descriptions

Validation choice task on the first wave

On the first wave of the study, the validation task labeled **GiftC** is expressed as:

Assuming you have received a store gift certificate from a relative that gives you enough money to pay for two products shown in the following list. One day after work, you go on a shopping trip to the mall. You see the following ten products, the same as the ones you evaluated previously, what purchase decisions would you make? Please select two products you will actually buy.

Validation choice tasks on the second wave

On the second wave of the study, I address the validation questions in two different ways.

Task 1 **GiftC** is the same as the one on the first wave:

Assuming you have received a store gift certificate from a relative that gives you enough money to pay for two products shown in the following list. One day after work, you go on a shopping trip to the mall. You see the following ten products, the same as the ones you evaluated previously, what purchase decisions would you make? Please select two products you will actually buy.

And Task 2 labeled **GiftR** asked:

Assuming we are offering a free gift registry service to make it easy for your family and friends to choose the gifts that you need and really want. We offer ten items for you to choose to put on your gift registry. These are the 10 products you evaluated a moment ago. If you are interested in getting any of these ten products as a gift, please register it by checking the box over the product picture.

Validation choice tasks on the third wave

On the third wave of the study, the two validations tasks are expressed almost the same as the second wave except I include an $1800 financial limit for the task 1.

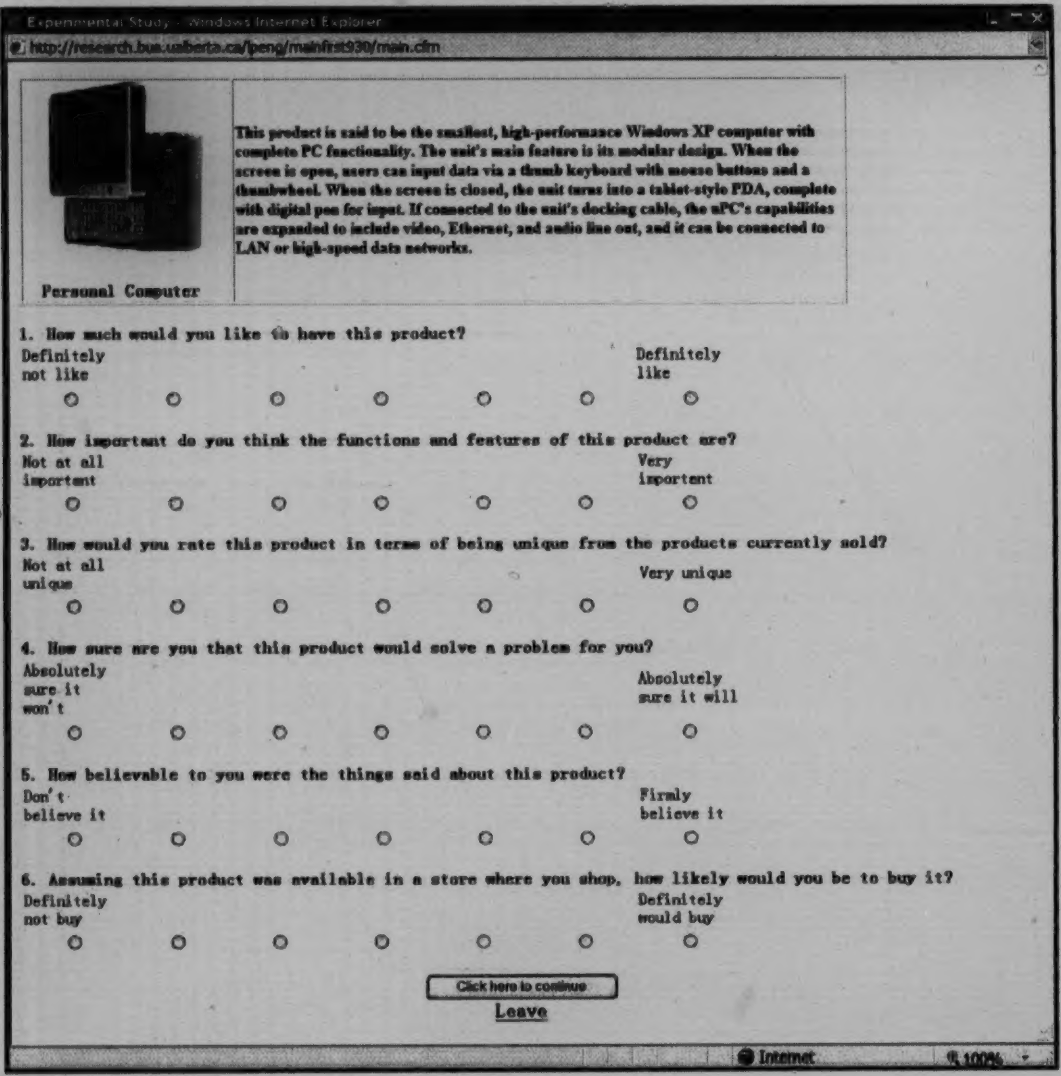
The task 1 labeled **GiftCM** asked:

Assuming you have received a store gift certificate from a relative that gives you $1800 that is only valid for the products shown in the following list. One day after work, you go on a shopping trip to the store. You see the following ten products with the retail prices shown over the pictures (the same as the ones you evaluated previously). What purchase decisions would you make out of your gift certificate? Please identify each of the products you would buy.

Task 2 labeled **GiftR** asked:

Assuming we are offering a free gift registry service to make it easy for your family and friends to choose the gifts that you need and really want. We offer ten items with the retail price over the picture for you to choose to put on your gift registry. These are the 10 products you evaluated a moment ago. If you are interested in getting any of these ten products as gifts that would be paid for by friends or relatives, please register them by checking the box over the product pictures.

8. An Example of Concept Evaluation Response Screen


Experimental Study - Windows Internet Explorer
http://research.bus.ualberta.ca/lpeng/mainfirst930/main.cfm
Click here to continue
Leave
Internet
100%

Personal Computer

This product is said to be the smallest, high-performance Windows XP computer with complete PC functionality. The unit's main feature is its modular design. When the screen is open, users can input data via a thumb keyboard with mouse buttons and a thumbwheel. When the screen is closed, the unit turns into a tablet-style PDA, complete with digital pen for input. If connected to the unit's docking cable, the uPC's capabilities are expanded to include video, Ethernet, and audio line out, and it can be connected to LAN or high-speed data networks.

1. How much would you like to have this product?

Definitely not like ○ ○ ○ ○ ○ ○ ○ Definitely like

2. How important do you think the functions and features of this product are?

Not at all important ○ ○ ○ ○ ○ ○ ○ Very important

3. How would you rate this product in terms of being unique from the products currently sold?

Not at all unique ○ ○ ○ ○ ○ ○ ○ Very unique

4. How sure are you that this product would solve a problem for you?

Absolutely sure it won't ○ ○ ○ ○ ○ ○ ○ Absolutely sure it will

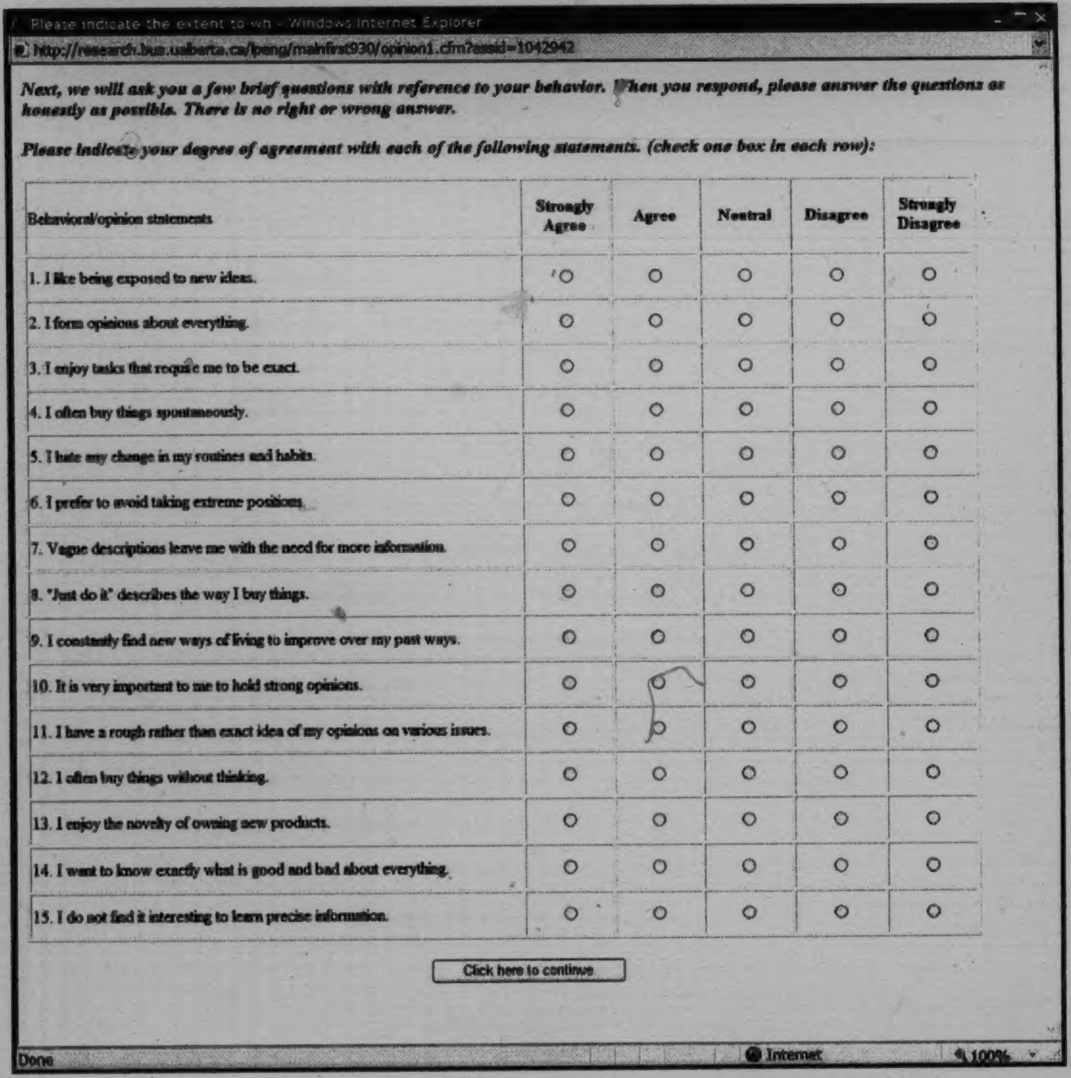
5. How believable to you were the things said about this product?

Don't believe it ○ ○ ○ ○ ○ ○ ○ Firmly believe it

6. Assuming this product was available in a store where you shop, how likely would you be to buy it?

Definitely not buy ○ ○ ○ ○ ○ ○ ○ Definitely would buy

9. An Example of Personality Trait Measures Response Screen

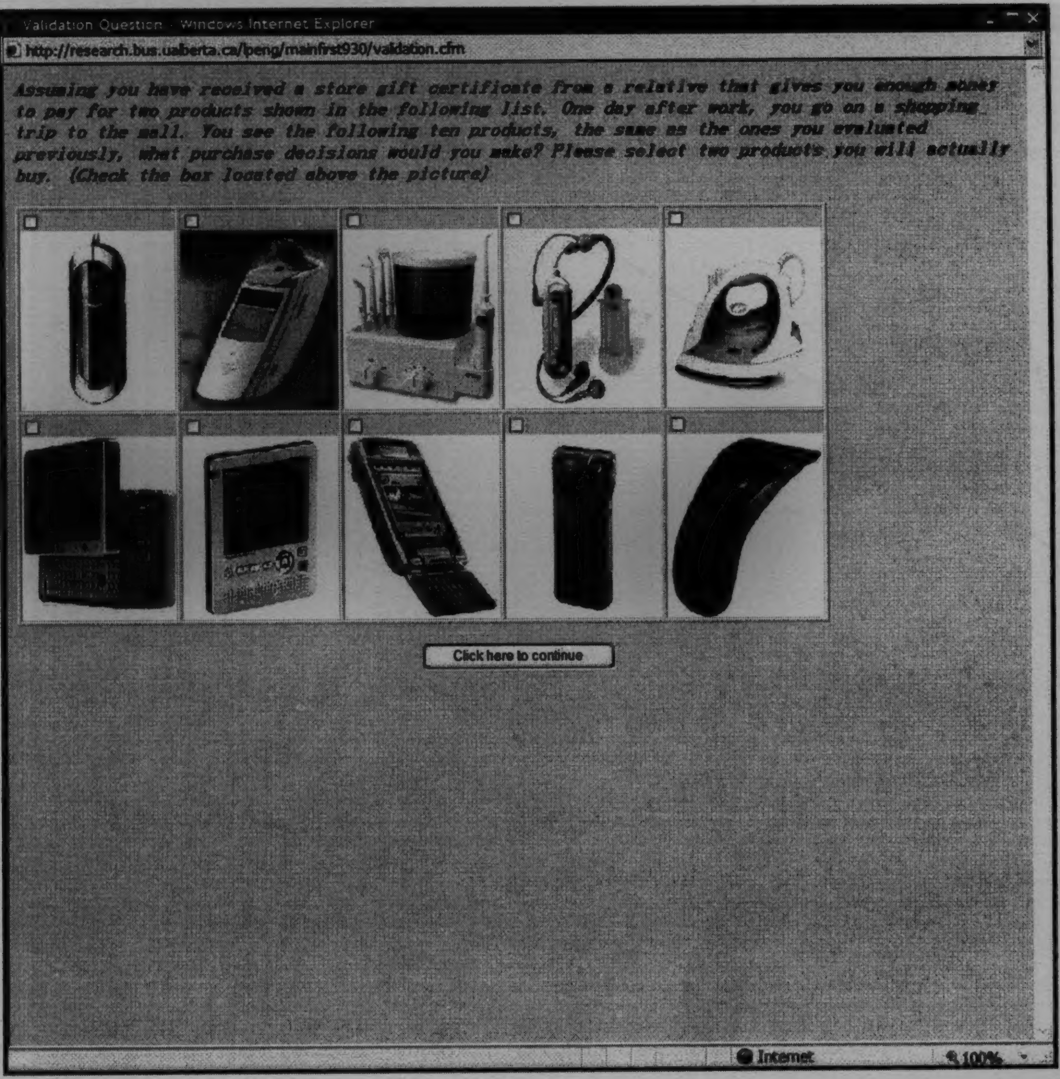

Please indicate the extent to wh - Windows Internet Explorer
http://research.bus.ualberta.ca/peng/mainfirst930/opinion1.cfm?assid=1042942
Click here to continue
Done
Internet
100%

Next, we will ask you a few brief questions with reference to your behavior. When you respond, please answer the questions as honestly as possible. There is no right or wrong answer.

Please indicate your degree of agreement with each of the following statements. (check one box in each row):

Behavioral/opinion statements	**Strongly Agree**	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**
1. I like being exposed to new ideas.	○	○	○	○	○
2. I form opinions about everything.	○	○	○	○	○
3. I enjoy tasks that require me to be exact.	○	○	○	○	○
4. I often buy things spontaneously.	○	○	○	○	○
5. I hate any change in my routines and habits.	○	○	○	○	○
6. I prefer to avoid taking extreme positions.	○	○	○	○	○
7. Vague descriptions leave me with the need for more information.	○	○	○	○	○
8. "Just do it" describes the way I buy things.	○	○	○	○	○
9. I constantly find new ways of living to improve over my past ways.	○	○	○	○	○
10. It is very important to me to hold strong opinions.	○	○	○	○	○
11. I have a rough rather than exact idea of my opinions on various issues.	○	○	○	○	○
12. I often buy things without thinking.	○	○	○	○	○
13. I enjoy the novelty of owning new products.	○	○	○	○	○
14. I want to know exactly what is good and bad about everything.	○	○	○	○	○
15. I do not find it interesting to learn precise information.	○	○	○	○	○

10. An Example of Validation Choice Tasks Response Screen


Validation Question - Windows Internet Explorer
http://research.bus.ualberta.ca/lpeng/mainfirst930/validation.cfm
Click here to continue
Internet
100%

Assuming you have received a store gift certificate from a relative that gives you enough money to pay for two products shown in the following list. One day after work, you go on a shopping trip to the mall. You see the following ten products, the same as the ones you evaluated previously, what purchase decisions would you make? Please select two products you will actually buy. (Check the box located above the picture)

A-5 Output from SPSS GLM - Repeated Measures Analysis

1. Mean differences among occasions

Within-Subjects Factors

Measure: SCORE

TIME	Dependent Variable
1	TIME1
2	TIME2
3	TIME3

Descriptive Statistics

	Mean	Std. Deviation	N
TIME1	5.1645	1.74351	4680
TIME2	4.8726	1.75820	4680
TIME3	4.8615	1.81633	4680

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
TIME	Pillai's Trace	.042	103.144[a]	2.000	4678.000	.000
	Wilks' Lambda	.958	103.144[a]	2.000	4678.000	.000
	Hotelling's Trace	.044	103.144[a]	2.000	4678.000	.000
	Roy's Largest Root	.044	103.144[a]	2.000	4678.000	.000

a. Exact statistic

b.

Design: Intercept
Within Subjects Design: TIME

Mauchly's Test of Sphericity[b]

Measure: SCORE

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.983	81.238	2	.000	.983	.983	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.

Design: Intercept
Within Subjects Design: TIME

Tests of Within-Subjects Effects

Measure: SCORE

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	276.309	2	138.155	116.469	.000
	Greenhouse-Geisser	276.309	1.966	140.533	116.469	.000
	Huynh-Feldt	276.309	1.967	140.475	116.469	.000
	Lower-bound	276.309	1.000	276.309	116.469	.000
Error(TIME)	Sphericity Assumed	11100.357	9358	1.186		
	Greenhouse-Geisser	11100.357	9199.619	1.207		
	Huynh-Feldt	11100.357	9203.453	1.206		
	Lower-bound	11100.357	4679.000	2.372		

Tests of Within-Subjects Contrasts

Measure: SCORE

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Linear	214.821	1	214.821	166.218	.000
	Quadratic	61.488	1	61.488	56.935	.000
Error(TIME)	Linear	6047.179	4679	1.292		
	Quadratic	5053.178	4679	1.080		

Tests of Between-Subjects Effects

Measure: SCORE

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	346276.003	1	346276.003	49063.05	.000
Error	33023.331	4679	7.058		

2. Mean differences between major innovations and minor innovations

Within-Subjects Factors

Measure: SCORE

TIME	Dependent Variable
1	TIME1
2	TIME2
3	TIME3

Between-Subjects Factors

		N
NEWNESS	1.00	2340
	2.00	2340

Descriptive Statistics

	NEWNESS	Mean	Std. Deviation	N
TIME1	1.00	5.0679	1.75256	2340
	2.00	5.2611	1.72939	2340
	Total	5.1645	1.74351	4680
TIME2	1.00	4.8026	1.75653	2340
	2.00	4.9427	1.75745	2340
	Total	4.8726	1.75820	4680
TIME3	1.00	4.8030	1.80827	2340
	2.00	4.9201	1.82286	2340
	Total	4.8615	1.81633	4680

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
TIME	Pillai's Trace	.042	103.175[a]	2.000	4677.000	.000
	Wilks' Lambda	.958	103.175[a]	2.000	4677.000	.000
	Hotelling's Trace	.044	103.175[a]	2.000	4677.000	.000
	Roy's Largest Root	.044	103.175[a]	2.000	4677.000	.000
TIME * NEWNESS	Pillai's Trace	.001	1.340[a]	2.000	4677.000	.262
	Wilks' Lambda	.999	1.340[a]	2.000	4677.000	.262
	Hotelling's Trace	.001	1.340[a]	2.000	4677.000	.262
	Roy's Largest Root	.001	1.340[a]	2.000	4677.000	.262

a. Exact statistic

b.

Design: Intercept+NEWNESS
Within Subjects Design: TIME

Mauchly's Test of Sphericity[b]

Measure: SCORE

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.983	80.900	2	.000	.983	.984	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.

Design: Intercept+NEWNESS
Within Subjects Design: TIME

Tests of Within-Subjects Effects

Measure: SCORE

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	276.309	2	138.155	116.482	.000
	Greenhouse-Geisser	276.309	1.966	140.524	116.482	.000
	Huynh-Feldt	276.309	1.968	140.435	116.482	.000
	Lower-bound	276.309	1.000	276.309	116.482	.000
TIME * NEWNESS	Sphericity Assumed	3.560	2	1.780	1.501	.223
	Greenhouse-Geisser	3.560	1.966	1.810	1.501	.223
	Huynh-Feldt	3.560	1.968	1.809	1.501	.223
	Lower-bound	3.560	1.000	3.560	1.501	.221
Error(TIME)	Sphericity Assumed	11096.798	9356	1.186		
	Greenhouse-Geisser	11096.798	9198.261	1.206		
	Huynh-Feldt	11096.798	9204.062	1.206		
	Lower-bound	11096.798	4678.000	2.372		

Tests of Within-Subjects Contrasts

Measure: SCORE

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Linear	214.821	1	214.821	166.275	.000
	Quadratic	61.488	1	61.488	56.925	.000
TIME * NEWNESS	Linear	3.385	1	3.385	2.620	.106
	Quadratic	.175	1	.175	.162	.688
Error(TIME)	Linear	6043.794	4678	1.292		
	Quadratic	5053.004	4678	1.080		

Tests of Between-Subjects Effects

Measure: SCORE
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	346276.003	1	346276.003	49170.38	.000
NEWNESS	79.125	1	79.125	11.236	.001
Error	32944.206	4678	7.042		

A-6 Confidence Intervals Mat Lab Calculation Code

Many applications in a variety of fields require the measurement of variance. To effectively understand these measurements, decision makers require both point and interval estimates. The book of Burdick and Graybill (1992) contains methods for constructing confidence intervals on individual variance components, linear combinations of variance components, and ratios of variance components for a variety of experimental designs. These designs include both crossed and nested factors, with both balanced and unbalanced data sets.

As shown in Burdick and Graybill (1992), the steps involved in constructing confidence intervals for a variance component are:

1. Express a variance component as linear combinations or ratio and linear combinations of EMS.
2. Fit the expression of variance component estimate to the following situations:
    - Single EMS, using (3.2.1)
    - Sums of EMS, using (3.2.5)
    - Linear combinations of EMS, using (3.3.3) and (3.3.4)
    - Ratio of EMS, using (3.4.1)
    - Combinations of the sum of EMS and the ratio of EMS, using (3.4.3) and (3.4.4)

Here I include the calculation code designed to construct confidence intervals for variance components. Such a variance component estimate can be expressed as linear combinations of Expected Mean Squares (EMS) with different signs. Please refer to the detailed equation (3.3.3) and (3.3.4) in Burdick and Graybill (1992).

```
Function CI(inputfile)
MAXVALUE = 10^8;

fid = fopen(inputfile, 'r');
outfile = sprintf('%s.res', inputfile);
fidout = fopen(outfile, 'w');
alpha = 1 - 0.05;
        while(~feof(fid))
        firstline = fgetl(fid);
        [t m] = size(firstline);
        if(m <= 0)
                continue;
        end;
        if(firstline(1) ~= '%')
                continue;
        end;
        if(firstline(2) == 'F')
                continue;
        end;

        fprintf(fidout,'%s\n', firstline);
```

```
    P = fscanf(fid, '%d',1);
    Q = fscanf(fid, '%d',1);
    Nq = fscanf(fid, '%f', P);
    Cq = fscanf(fid, '%f', P);
    Sq = fscanf(fid, '%f', P);
    Nr = fscanf(fid, '%f', Q-P);
    Cr = fscanf(fid, '%f', Q-P);
    Sr = fscanf(fid, '%f', Q-P);
    Nq = Nq';
    Cq = Cq';
    Sq = Sq';
    Nr = Nr';
    Cr = Cr';
    Sr = Sr';
    Gq = 0;

%Computing the lower bound;

    Delta = sum(Cq.*Sq) - sum(Cr.*Sr);
    for i = 1:P
        Gq(i) = 1 - 1/finv(alpha, Nq(i), MAXVALUE);
    end
    for i = 1:Q-P
        Hr(i) = 1/finv(1-alpha, Nr(i), MAXVALUE)-1;
    end
    for i = 1:P
        for j = 1:Q-P
            Gqr(i,j) = ((finv(alpha, Nq(i), Nr(j)) - 1)^2 -
            Gq(i)^2*finv(alpha,Nq(i),Nr(j))^2 - Hr(j)^2)/finv(alpha,
            Nq(i),Nr(j));
        end;
    end;
    if(P == 1)
        Gqt(1,1) = 0;
    else
        for i = 1:P
            for j = i+1:P
                Gqt(i,j) = ((1-1/finv(alpha,Nq(i)+Nq(j),MAXVALUE))^2
                *(Nq(i)+Nq(j))^2/(Nq(i)*Nq(j)) - Gq(i)^2*Nq(i)/Nq(j) -
                Gq(j)^2*Nq(j)/Nq(i))/(P-1);
            end;
        end;
    end;
    Vl1 = sum(Gq.^2.*Cq.^2.*Sq.^2);
    Vl2 = sum(Hr.^2.*Cr.^2.*Sr.^2);
    Vl3 = 0.0;
    for i = 1:P
        for j = 1:Q-P
            Vl3 = Vl3 + Gqr(i,j)*Cq(i)*Cr(j)*Sq(i)*Sr(j);
        end;
    end;

    Vl4 = 0.0;
    for i = 1:P-1
        for j = i+1:P
            Vl4 = Vl4 + Gqt(i,j)*Cq(i)*Cq(j)*Sq(i)*Sq(j);
        end;
    end;
    Vl = Vl1 + Vl2 + Vl3 + Vl4;
    if(Vl < 0)
        Vl = 0;
    end;
    L = Delta - sqrt(Vl);


%Computing the upper bound;
    for i = 1:P
        Hq(i) = 1/finv(1-alpha, Nq(i), MAXVALUE) - 1;
    end;
    for i = 1:Q-P
        Gr(i) = 1 - 1/finv(alpha, Nr(i), MAXVALUE);
    end;
    for i = 1:P
        for j = 1:Q-P
            Hqr(i,j) = ((1-finv(1-alpha, Nq(i), Nr(j)))^2 - Hq(i)^2*finv(1-
            alpha,Nq(i),Nr(j))^2 - Gr(j)^2)/finv(1-alpha, Nq(i),Nr(j));
        end;
    end;
    if(Q-P-1 == 0)
        Hru(1,1) = 0;
    else
        for i = 1:Q-P
            for j = i+1:Q-P
                Hru(i,j) = ((1-1/finv(alpha,Nr(i)+Nr(j),MAXVALUE))^2
                *(Nr(i)+Nr(j))^2/(Nr(i)*Nr(j)) - Gr(i)^2*Nr(i)/Nr(j) -
                Gr(j)^2*Nr(j)/Nr(i))/(Q-P-1);
            end;
        end;
    end;
    Vu1 = sum(Hq.^2.*Cq.^2.*Sq.^2);
    Vu2 = sum(Gr.^2.*Cr.^2.*Sr.^2);
    Vu3 = 0.0;
```

```
          for i = 1:P
                    for j = 1:Q-P
                              Vu3 = Vu3 + Hqr(i,j)*Cq(i)*Cr(j)*Sq(i)*Sr(j);
                    end;
          end;
          Vu4 = 0.0;
          for i = 1:Q-P-1
                    for j = i+1:Q-P
                              Vu4 = Vu4 + Hru(i,j)*Cr(i)*Cr(j)*Sr(i)*Sr(j);
                    end;
          end;
          Vu = Vu1 + Vu2 + Vu3 + Vu4;
          if(Vu < 0)
                    Vu = 0;
          end;
          U = Delta + sqrt(Vu);
          fprintf(fidout, 'Upperbound: %f', U);
          fprintf(fidout, '    Lowerbound: %f\n', L);
          end;
fclose(fidout);
fclose(fid);
return;
```

A-7 Output from SPSS Factor Analysis for the Observed Trait Data

Pattern Matrix

	Factor			
	1	2	3	4
NES	-.111	.300	-.111	.683
NFP	.170	-.244	3.542E-02	.947
IMPUL	-.101	.461	-2.333E-02	-4.243E-02
DSI	.334	.548	-.317	8.785E-02
CIJM	-.100	.667	.111	1.582E-02
CNS	.607	.537	-.148	2.778E-02
DUCP	.259	.678	-4.385E-02	.127
SD	.489	-3.401E-02	-1.944E-03	9.422E-02
INVOL	.626	-.140	5.641E-02	2.608E-02
CSO	5.598E-02	.197	-.678	5.128E-02
PFC	.212	.181	.838	-.103
EAP	.382	2.478E-03	-.772	-7.234E-02
EIS	.490	.225	-.272	5.472E-02

Extraction Method: Principal Axis Factoring. Rotation Method: Oblimin with Kaiser Normalization. a Rotation converged in 12 iterations.

Factor Correlation Matrix

Factor	1	2	3	4
1	1.000	.116	-.291	.311
2	.116	1.000	-.257	.126
3	-.291	-.257	1.000	-.369
4	.311	.126	-.369	1.000

Extraction Method: Principal Axis Factoring. Rotation Method: Oblimin with Kaiser Normalization.

Communalities

	Initial	Extraction
NES	.646	.643
NFP	.648	.995
IMPUL	.319	.216
DSI	.795	.766
CIJM	.417	.421
CNS	.825	.866
DUCP	.731	.654
SD	.341	.274
INVOL	.303	.386
CSO	.702	.627
PFC	.565	.663
EAP	.761	.861
EIS	.637	.533

Extraction Method: Principal Axis Factoring.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.957	38.128	38.128	4.662	35.864	35.864	2.586
2	1.850	14.232	52.359	1.435	11.036	46.899	2.536
3	1.395	10.730	63.089	.958	7.370	54.270	3.140
4	1.088	8.372	71.461	.849	6.527	60.797	2.334
5	.942	7.248	78.709				
6	.724	5.567	84.277				
7	.541	4.158	88.435				
8	.468	3.598	92.033				
9	.329	2.531	94.564				
10	.311	2.394	96.958				
11	.165	1.269	98.227				
12	.128	.984	99.211				
13	.103	.789	100.000				

Extraction Method: Principal Axis Factoring.
a When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

A-8 Output from SPSS Mixed Model Analysis for DSI

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters
Fixed Effects	Intercept	1		1
	DSIMEAN	1		1
Random Effects	ITEM	6	Identity	1
	PERSON	78	Identity	1
	TIME	3	Identity	1
	CONCEPT	10	Identity	1
Residual				1
Total		99		7

a. Dependent Variable: SCORE.

Information Criteria[a]

-2 Restricted Log Likelihood	50234.88
Akaike's Information Criterion (AIC)	50244.88
Hurvich and Tsai's Criterion (AICC)	50244.88
Bozdogan's Criterion (CAIC)	50287.63
Schwarz's Bayesian Criterion (BIC)	50282.63

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: SCORE.

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	74.320	53.924	.000
DSIMEAN	1	76.000	10.720	.002

a. Dependent Variable: SCORE.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	3.591657	.4891060	74.320	7.343	.000	2.6171621	4.5661525
DSIMEAN	.4240636	.1295181	76.000	3.274	.002	.1661060	.6820212

a. Dependent Variable: SCORE.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Residual		2.036524	2.44E-02
ITEM	ID diagonal	.2148530	.1364354
PERSON	ID diagonal	.7730521	.1272411
TIME	ID diagonal	2.91E-02	2.95E-02
CONCEPT	ID diagonal	7.41E-02	3.56E-02

a. Dependent Variable: SCORE.

A-9 Venn Diagrams for the G Study Designs

"In a Venn diagram, each main effect is represented by a circle. Interaction effects are represented by the intersections of circles. The total number of effects in any design is the number of distinct areas in the Venn diagram. When a main effect involves nesting, it is represented by a circle within another circle, or within the intersection of circles." (Brennan 2001, p. 54-55)

Figure A-1 Venn diagram for the crsi design in Chapter 4

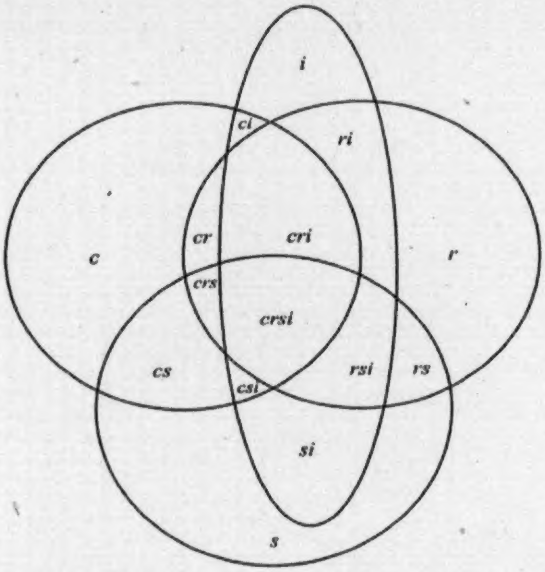


Figure A-2 Venn diagram for the G study design in academic data

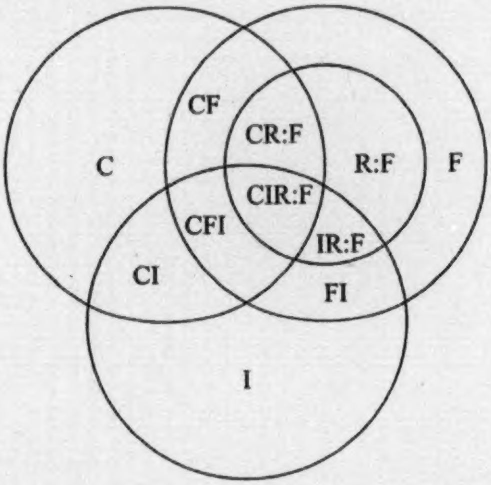


Figure A-3 Venn diagram for the G study design in FMCG data

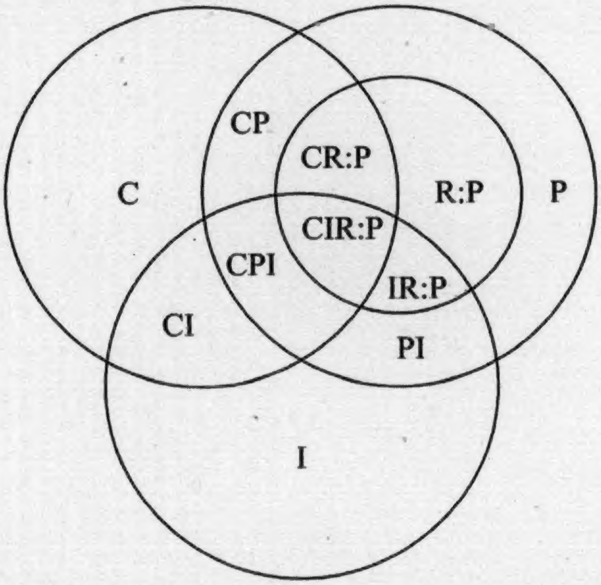


Figure A-4 Venn diagram for the G study design in Innovative data I

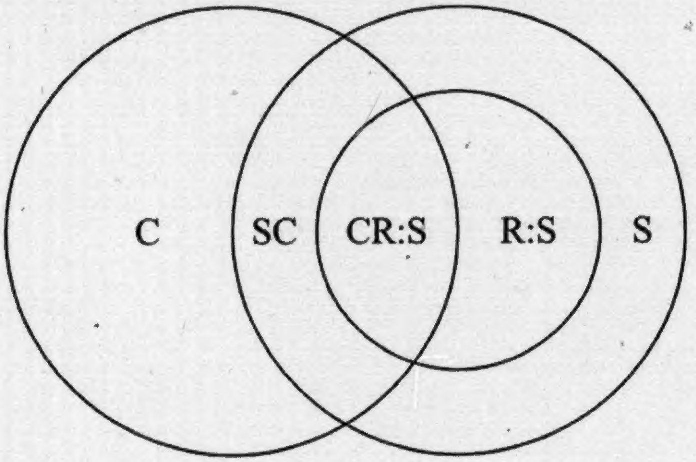


Figure A-5 Venn diagram for the G study design in Innovative data II

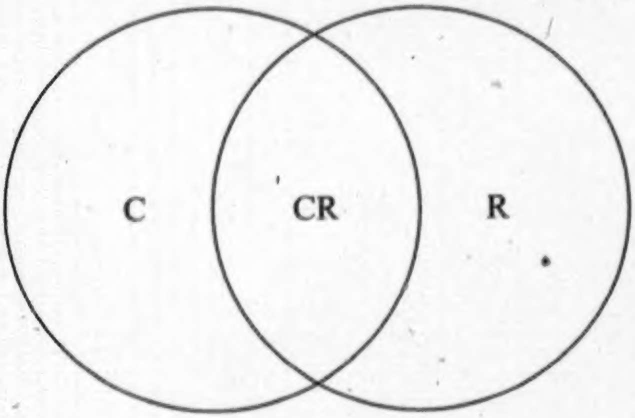


Figure A-6 Venn diagram for the G study design in the pretest

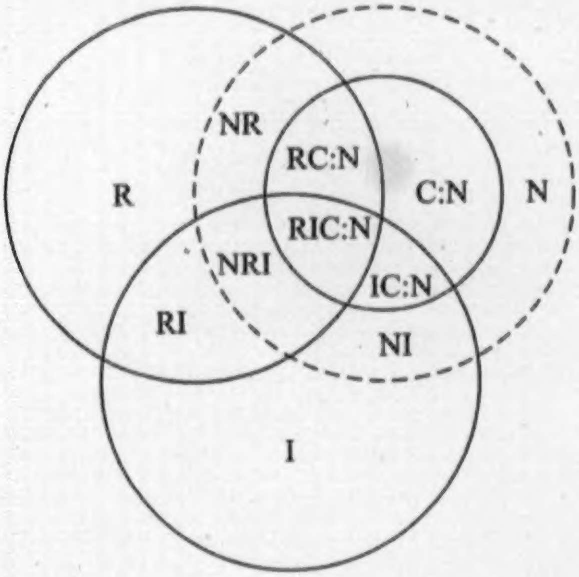


Figure A-7 Venn diagram for the G study design in Table 6-5

(Occasions as an explicit facet)

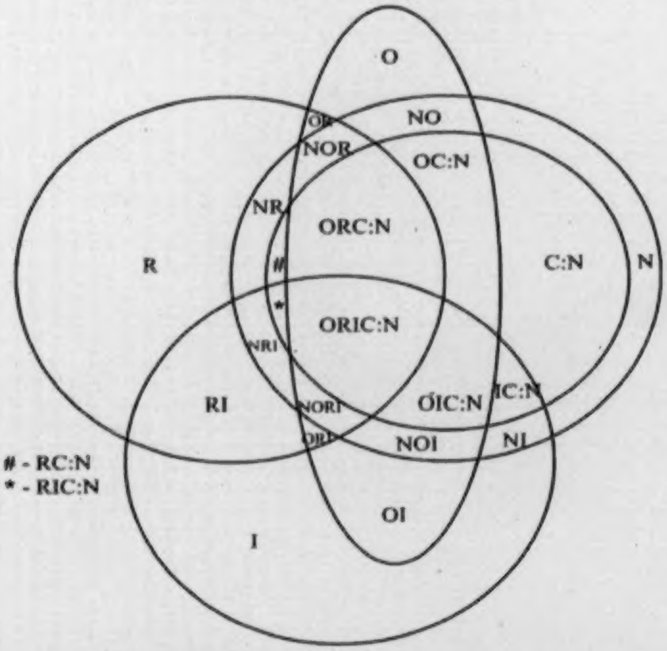


Figure A-8 Venn diagram for the G study design in Table 6-14

(Design issues revisited in the main study)

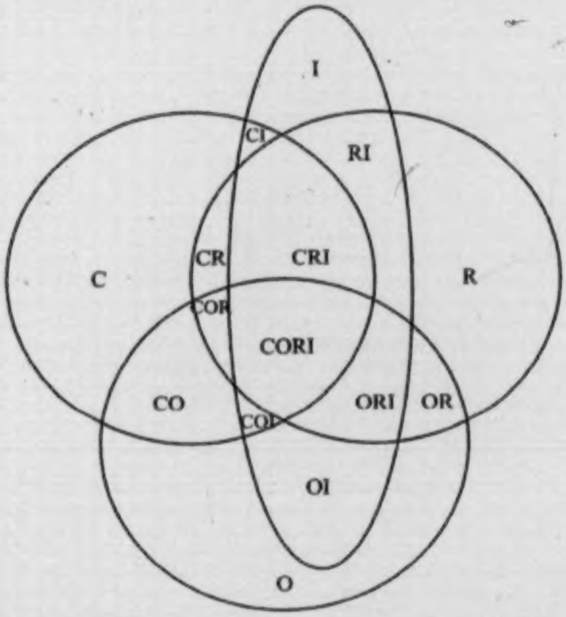


Figure 4-1
Conceptual Framework

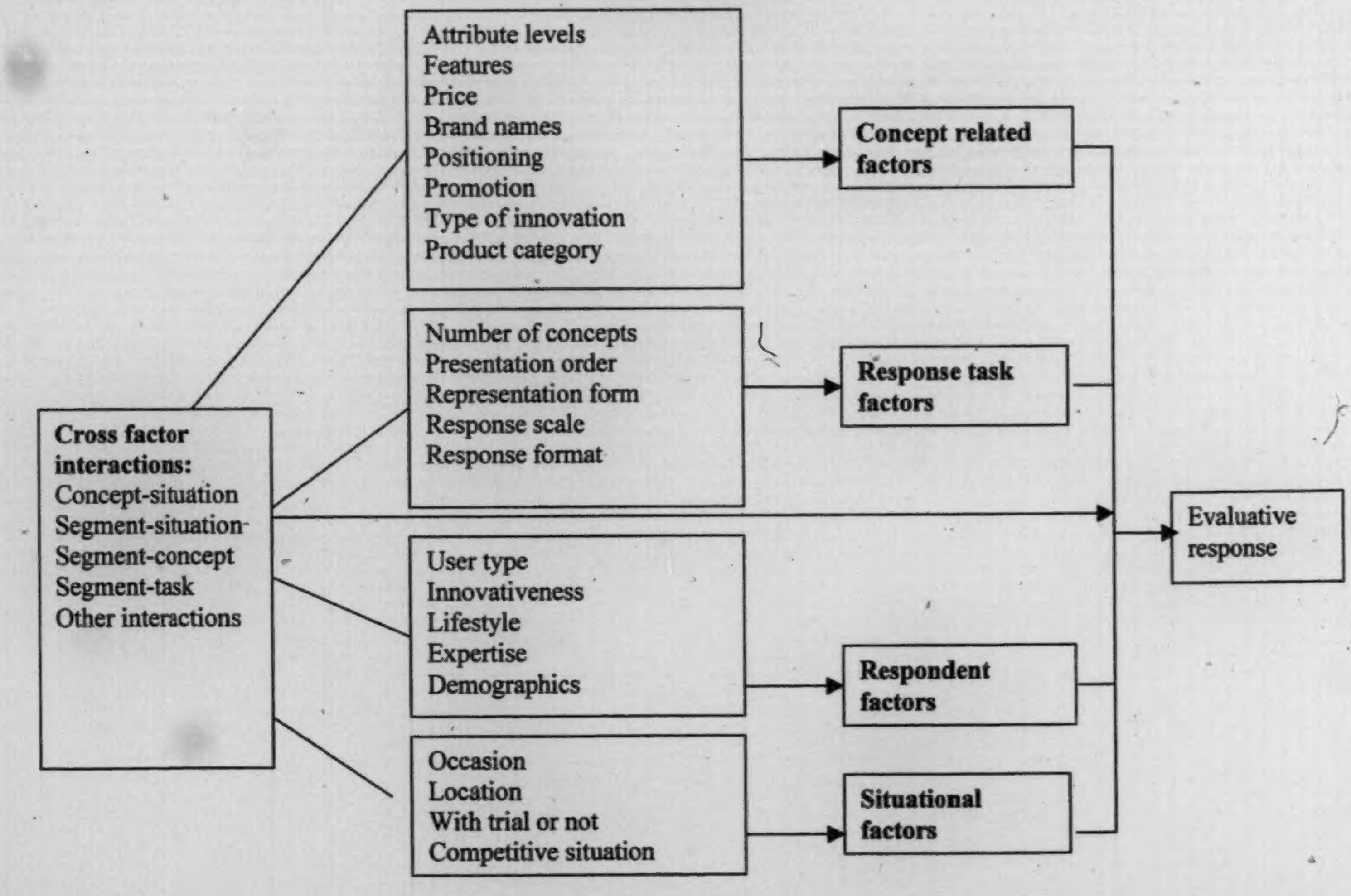


Table 2-1
Literature Review on Factors Influencing the Evaluative Scores

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Bengston & Brenner (1964)	Field experiment: Between subjects	Test methodologies	3 methods: side by side; staggered and monadic	Different test methodologies produce different test results; side by side magnify small differences, many of which may not be commercially important. A over B in side by side and about a stand-off in the staggered and monadic tests	Each technique has its place. Side by side most sensitive but most artificial; staggered is a comprise; monadic close to real life but insensitive
Armstrong & Overton (1971)	Field experiment: Within subjects	Methods of presentation: the extent and type of description of a new service	2 forms: Brief versus comprehensive description	The two descriptions were in substantial agreement in level of demand at various price and price elasticity and the identification of most and least likely users.	The extent and type of description did not seem to have an appreciable influence upon the results. The brief description proved to be much superior because of its low cost.
Haley & Gatty (1971)	Field experiment: Within subjects	Executions (Concept by copywriter	24 executions (8 different concepts by 3 copywriters)	The concepts received sharply differing ratings depending on which copywriter did the work.	Consumer reaction is based not only on the concept and the copywriter but also on the interaction between positioning and copywriter.

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Hughes & Guerrero (1971)	Computer-controlled Experiment: Within subjects	Sources of information, Copy themes and product dimensions	3 levels of sources: social, neutral and commercial 3 levels of themes: rating of 20, 50 and 80 3 dimensions: performance, comfort and safety, social acceptability	Significant effects for the elements of source, theme and product dimension; Significant interactions for sources-themes, theme-dimensions; The theme of message (concepts) is the dominating influence in this experiment.	Simultaneous concept testing offer one means for reducing cost of concept tests. It yields results that are statistically significant and reliable. It is ready for validation test in the field.
Tauber (1972)	Field experiment: Between subjects	Communication form	2 forms: A proto-typical print ad with pictorial stimuli versus a factual written description	Scores (overall attitude and intention to buy) were much higher for the print ad than the paragraphs but the relative scores did not change.	Thus the one big idea come through in either form of communication
Wolpert (1980)	Review	Respondent segments	2 segments: common car buyers versus nonfunctional show type styling buyers	The former segment prefers a conservative functional styling with little change; the latter may be able to give valid predictions of the acceptance of revolutionary change in styling	Innovators and early adopters' reactions to concepts may serve as leading indicators of the acceptance of major innovations.

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Holbrook & Moore (1981)	Lab experiment: Conjoint study; Between subjects	Representation form	2 forms: schematic pictorial representations and verbal descriptions	Pictures evoked significantly more main effects than did words. No significant difference in the number of attribute interactions was found.	For fashion-oriented products, pictorial displays elicit a greater number of main feature effects than does verbal descriptions. Individual cognitive strategies moderate these effects.
Smead et al. (1981)	Lab experiment: Conjoint study; Between subjects	Representation forms in choice experiments	2 forms: actual products and verbal representations	Choices from verbal representations were perceived to be easier than choices from actual products; actual products elicited more scan behavior (eye movements) than did verbal; and there were differences in the determinant attributes between two modes.	Researchers should be wary of producing descriptions that are more easily processed than real products or that distort the use of different kinds of attributes.
Wilton & Pessemier (1981)	Field experiment: Conjoint study; Between subjects	The amount and importance of the information given to consumers; Purchase context	3 information levels: low, intermediate and maximum; Control group: no information; 2 purchase contexts: low vs. high risk	Increased levels of information have an effect on the predicted market share of an electric car; Information levels has a significant effect on perceptions and preferences for the products	How a sponsor communicates about an innovative product or service can influence perceptions, preferences, and choice.

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Moore & Holbrook (1982)	Lab experiment: Conjoint study; Within subjects	Representation form	2 forms: Real product versus concepts	The shift from real objects to new concepts is accompanied by a dramatic decline in predictive efficacy in both joint space models and conjoint analysis.	The perceptions of consumers actually differ when they see a concept or a real product because the affective overtones increase predictive fits between preference and perceptions.
Trebbi & Flesch (1983)	Field experiment: Between subjects	The number of concepts being tested	2 levels: Monadic and sequential monadic (multiple)	Concepts were rated significantly lower in the multiple treatment in terms of the evaluative variable (purchase likelihood); a greater level of concurrence on the perceptual variables (perceived novelty and performance confidence)	Single and multiple testing contexts yield substantially different retention results. These results speak strongly against the advisability of capitalizing upon the cost savings afforded by testing concepts in a multiple context.
Lewis (1984)	Case study	Concept tests with product placement and without product placement	2 levels: with product trial or without	Concept tests with product placement results in higher positive interest in concepts	Consumers respond to the method rather than the message

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Reidenbach & Grimes (1984)	Lab experiment: Between subjects	The type of innovation The way the concept is presented Knowledge level of respondents	2 concepts: continuous and discontinuous innovation 3 presentation formats: written, verbal and video 2 knowledge level: high and low	High knowledge groups evaluate concepts differently than low knowledge groups. This differential impact is conditioned by the type of concept being evaluated and the form by which the concept is presented	High knowledge groups provide more accurate evaluations
Domzal & Unger (1985)	Lab experiment: Between subjects	Representation form	2 forms: schematic pictorial and verbal 3 features: leather versus metal band; round versus square face; digital versus analog function	No significant differences in the number of significant main attribute effects. Pictorial presentation of stimuli generated significantly fewer interaction effects than did verbal presentation.	Some differences in consumer processing of verbal and pictorial information may be explained by the nature of the product (functional versus aesthetic) being evaluated.
Anderson (1987)	Experiment: conjoint study; Within subjects	Representation forms	3 levels: actual (unlabeled) products; actual products with verbal descriptions (labeled products) and verbal representations only	Obtained the highest fit (R2) values under verbal representations followed by labeled products; the mean part-worths under verbal and unlabeled differed significantly, although the order of the part-worths within attributes was identical	

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Louviere et al. (1987)	Choice experiment: Between subjects	Stimulus presentation	2 levels: verbal descriptions versus partial realistic representations	Few differences in part-worths between representation modes	The research suggests reliance on visual wherever possible because considerable effort and expense is necessary to insure comparability of verbal and visual attribute descriptions.
Miller, et al. (1987)	Field experiment: Between subjects	Competitive-set information	2 levels: no competitive-set information provided versus competitive-set information provided 4 price levels	The provision of competitive-set information does induce higher purchase intentions, but not significant.	Competitive-set data can make a difference in products with luxury characters. Need assess the effects in different categories of products.
Schoormans et al. (1985)	Lab experiment: Between subjects	Consumer expertise	3 levels product-category expertise: low, moderate, high	High expertise consumers show more similarity between concept evaluation and the evaluation of the real product than consumers with little or moderate product-category expertise, produce more consistent and more stable evaluations over time.	The presence of product-category expertise enhances the ability of respondents to evaluate the concepts in a test for major and minor innovations.

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Dickinson & Wilby (1997)	Lab experiment: Between subjects	Positioning statement and product trial and their interrelationship	2 levels: product trial or absence of trial; 3 positioning: taste, gentleness and all natural ingredients	The test results did not reveal any interaction effects between product trial and product positioning. Main effects of positioning and product trial are significant	In the case of product line extensions for familiar consumer goods such as toothpaste, effective concept tests do not require product trials.
Vriens et al. (1998)	Field experiment: conjoint study; Within subjects	Representation forms	2 forms: verbal versus realistic pictorial 7 attributes manipulated	Design attributes have higher relative importances in the pictorial mode; A higher degree of respondent heterogeneity with pictorial mode; Verbal mode produced greater predictive accuracy.	The pictorial representations improved the respondents' understanding of the design attributes, while the verbal representations seem to facilitate judgment.
Dahan & Srinivasan (2000)	Lab experiment: Conjoint study; Between subjects	Representation form	4 forms: Attribute-only, full-profile; Web Static visual; Web Virtual animation; Physical Prototype	Virtual prototypes on the Web provide nearly the same market share predictions as physical prototypes for the bicycle pump product category.	The Web can help to reduce the uncertainty and cost of new product introductions by allowing more ideas to be concept tested in parallel.

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Kristensson et al. (2004)	Quasi-experiment: Between subjects	User type	3 types: Product Development professionals; advanced users and ordinary users	Ordinary users generate ideas that an independent panel judges as being more original and more valuable than the other two groups. On the other hand, advanced users and product development professionals develop ideas that are more realizable—more easily developed into actual products.	These findings provide interesting implications for whom to involve in idea generation, and when.
Dahl & Hoeffler (2004)	Lab experiment: Between subjects	Situational factor: visualization form	2 forms: Whether individuals are told to imagine themselves or someone else using a new product (self-related vs. others-related)	Evaluations are higher for incremental new products when self-visualization of the product is induced. However, product evaluations for radical new products are higher when an individual visualizes someone other than themselves using the new product.	Consumer-invoked visualization can have a significant effect on consumer reactions to new products
Lees & Wright (2004)	Field experiment: Between subjects	Concept formulation	3 forms: Stripped, embellished and visual	Respondents' answers to attitude and purchase intention questions showed only minor variation with different formulations. The ranking showed no substantial changes	Stripped concept statements are suggested for reducing costs and allowing organizations to undertake more frequently across a wider range of products

Authors & Year	Type of evidence	Factors Being Studied	No. Of fixed levels	Direction of Effect	Substantive conclusion
Creusen & Schoormans (2005)	Lab experiment and open-ended interview	The roles of product appearance	6 different roles: Have aesthetic and symbolic value; communicate functional characteristics; provide a quality impression; communicate ease of use; draw attention; influence the ease of product categorization.	Most subjects mentioned two different ways in which appearance influenced their product choice. The aesthetic and symbolic roles were mentioned most often.	Distinguishing these six appearance roles will help product development managers to optimize the product appearance better to market needs
Ziamou & Veryzer (2005)	Lab experiment: Between subjects	Temporal distance from the purchase or use occasion	2 occasions: Near versus distant future (tomorrow & a year from now; introduce in a few months & currently available)	The functionality of the product is valued more in distant future events while the interface of the product is more important in the near future.	This study suggests that the measures of consumer purchase intention in concept tests may be misleading.